Statistics 133 Final Project

Christian Alarcio, Ellen Chan, Anais Sidhu, Ruomeng (Michelle) Yang

December 7, 2015

Introduction

Our findings and analysis from this project were led by our hypothesis:

Technology industries show higher growth and higher beta, and therefore lower P/E ratios.

We answered the following questions:

- What is the relationship between beta and the payout percentage?
- What is the relationship between beta and expected growth rate?

Before we introduce our dataset and preprocessing steps, here is an explanation of some of the financial terminology.

- *PE ratio:* market value per share divided by earnings per share (EPS). It is a ratio for valuing a company through its current share price relative to its per-share earnings
- *Forward PE:* a measure of PE using forcasted earnings as a part of the claculation. Often called the "estimated price to earnings", it is calculated using the market price per share over expected earnings per share
- *Trailing PE:* the most commonly used PE measure and is based on actual earnings, and therefore more accurate. It is calculated by dividing current share price by the trailing twelve months' earnings per share
- *Beta:* a measure of volatility or risk of a company or industry in comparison to the market as a whole. There are two types of betas: levered and unlevered. The unlevered beta is the beta of a company without any debt or the measure of risk when removing the financial effects from adding debt to a firm's capital structure (finances). Levered beta is the beta of a company as a whole when accounting for debt
- *PEG:* a stock's PE ratio divided by the growth rate of its earnings for a specified time period. The ratio is used to determine a stock's value when taking into account a company's earnings growth and is considered to provide a more complete picture than PE
- *Payout ratio:* a proportion of earnings paid out as dividends to shareholders and is calculated by dividing divendens per share over earnings per share. It is known that the payout ratio is directly correlated with the PE ratio

Dataset & Preprocessing

We began by downloading two datasets as CSV files from the NYU Stern Business School's data archives. There was an online link that automatically downloaded the data we needed, so no R was needed. We then set the working directory and read the two CSV files. The files were then merged and assigned to the data frame "raw_data".

```r
# Set working directory
setwd("~/Documents/UC\ Berkeley\ 2015-2016/Statistics\ 133/projects/final/")

# Extract raw data
beta <- read.csv("raw_data/total_beta.csv", header = TRUE,
```

```
                  stringsAsFactors = FALSE)
pe <- read.csv("raw_data/pe_data.csv", header = TRUE, stringsAsFactors = FALSE)

# Combine data from the two files
raw_data <- merge(beta, pe, by = intersect(names(beta), names(pe)))
```

We then inspected the contents of the data frame.

```
# Inspect merged data
head(raw_data)
summary(raw_data)
names(raw_data)
str(raw_data)

# Inspect individual elements of the merged data
print("Summary of elements in raw_data")
for (i in 1:length(raw_data)) {
    print(paste0("Summary of ", names(raw_data)[i], ":"))
    print(summary(raw_data[, i]))
}
```

Further inspection revealed that column “X” is a column full of NA values. We start to clean data by removing this column by subsetting and reassigning this subset as the data frame “clean_data”. We also find that some columns have unideal names, which we rename.

```
# Duplicate raw_data to make edits
clean_data <- raw_data

# Remove blank columns
clean_data$X <- NULL

# Rename column titled Aggregate.Mkt.Cap..Net.Income
clean_data$Aggregate.Mkt.Cap.Net.Income <-
    clean_data$Aggregate.Mkt.Cap..Net.Income
clean_data$Aggregate.Mkt.Cap..Net.Income <- NULL

# Rename column titled clean_data$Aggregate.Mkt.Cap..Trailing.Net.Income
clean_data$Aggregate.Mkt.Cap.Trailing.Net.Income <-
    clean_data$Aggregate.Mkt.Cap..Trailing.Net.Income
clean_data$Aggregate.Mkt.Cap..Trailing.Net.Income <- NULL

# Rename column titled Number.of.firms
clean_data$Number.of.Firms <- clean_data$Number.of.firms
clean_data$Number.of.firms <- NULL
```

We then turn the two character columns, Average Correlation and Expected Growth in Next 5 Years, into numeric vectors to accurately analyze those numbers.

```
# Rename column titled Average.correlation and turn it into a numeric column
# vector
clean_data$Average.Correlation <-
    as.numeric(gsub("%", "", clean_data$Average.correlation))
clean_data$Average.correlation <- NULL
head(clean_data$Average.Correlation)
```

```
## [1] 16.05 30.81 31.63 19.57 22.34 25.75
```

```
# Rename column titled Expected.growth...next.5.years and turn it into a
# numeric column vector
clean_data$Expected.Growth.Next.5.Years <-
    as.numeric(gsub("%", "", clean_data$Expected.growth...next.5.years))
clean_data$Expected.growth...next.5.years <- NULL
head(clean_data$Expected.Growth.Next.5.Years)
```

```
## [1] 13.08 10.82 34.73 17.01 21.93 15.62
```

Then, we proceed to removing rows that contain data that cannot be evaluated or will not aid in our analysis. We found that the row "Unclassified" fit that description. We also considered removing "Total Market", since that did not pertain to the entire industry but is simply an aggregate of all other industries.

```
# Make copy of clean_data without the row "Unclassified"
clean_data <- clean_data[-which(clean_data$Industry == "Unclassified"),]

# Make copy of clean_data without the row "Unclassified" or "Total Market"
industries_only <-
    clean_data[-which(clean_data$Industry == "Unclassified" |
                          clean_data$Industry == "Total Market"),]
```

We then inspect elements of our cleaned dataset through finding the summaries, plots, min and max values, and histograms. For the complete code, see `code/preprocessing.R`. Finally, we create CSV files for our cleaned data and placed it into the clean_data directory.

```
# Create CSV files for clean data
file.create("clean_data/clean_data.csv")
write.csv(clean_data, file = "clean_data/clean_data.csv")
file.create("clean_data/industries_only.csv")
write.csv(industries_only, file = "clean_data/industries_only.csv")
```

Methods & Analysis

Before we can analyze our data, we will first need to set up the packages and dependencies we need. We will also need to retrieve and read our clean dataset before we begin. For our complete analysis, please see `code/analysis.R`, as some parts of the code are missing.

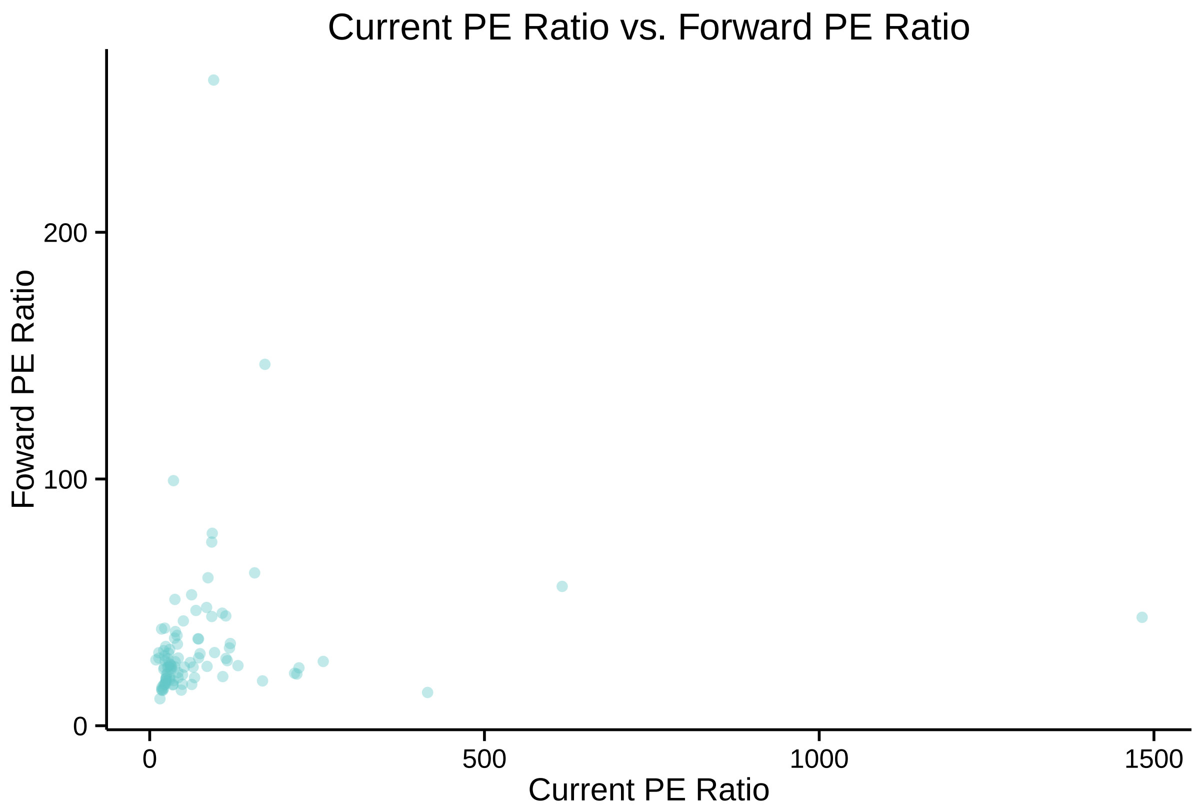
```
# Set working directory back for knitting
setwd("/Users/Michelle/Documents/UC Berkeley 2015-2016/Statistics 133/projects/final/report")

# Set up ggplot2
library(ggplot2)

# Set up readr
library(readr)

# Set up scatterplot3d
library(scatterplot3d)
```

```r
# Set up stringr
library(stringr)

# Set correct working directory again
setwd("~/Documents/UC\ Berkeley\ 2015-2016/Statistics\ 133/projects/final/")

# Read data files
clean_data <- read.csv("clean_data/clean_data.csv", header = TRUE,
                       row.names = 1, stringsAsFactors = FALSE)
industries_only <- read.csv("clean_data/industries_only.csv", header = TRUE,
                            row.names = 1, stringsAsFactors = FALSE)

# A look at clean_data
str(clean_data)
```

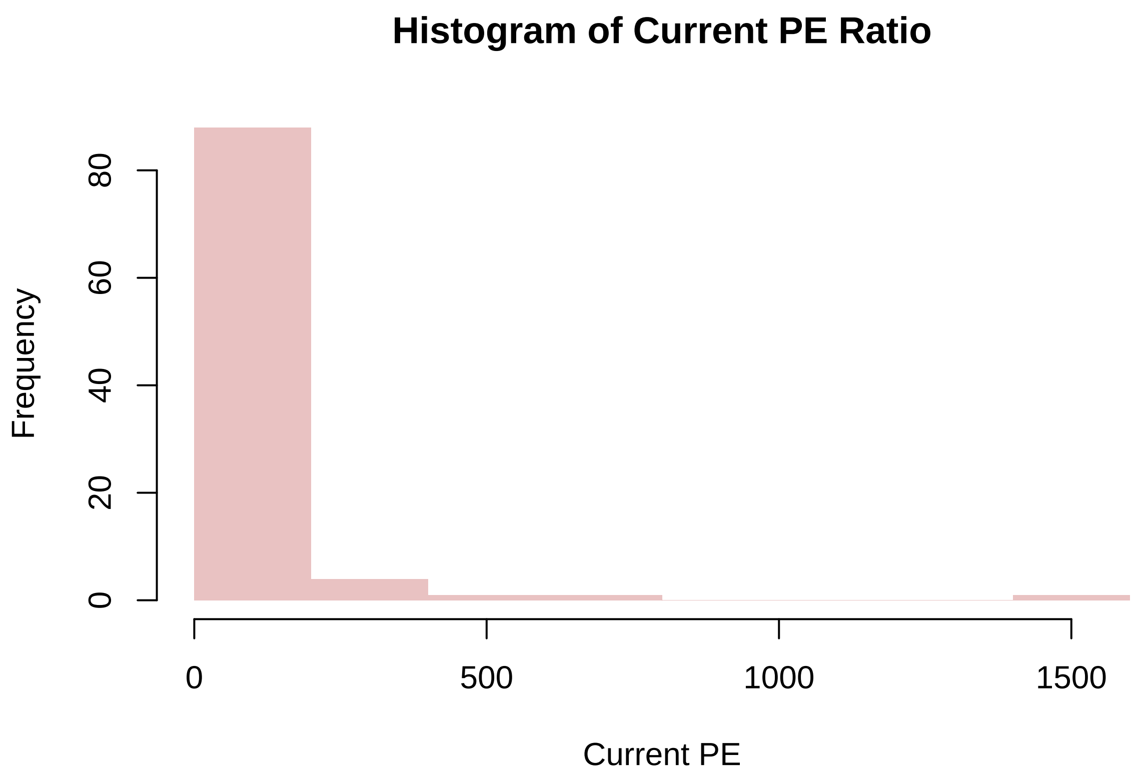
```
## 'data.frame':    95 obs. of  14 variables:
##  $ Industry                         : chr  "Advertising" "Aerospace/Defense" "Air Transport" "Ap
##  $ Average.Unlevered.Beta           : num  0.83 1.06 0.61 0.86 0.59 1.14 0.34 0.37 0.89 0.98 ...
##  $ Average.Levered.Beta             : num  1.18 1.16 0.98 0.99 1.09 1.35 0.81 0.53 1.06 1.14 ...
##  $ Total.Unlevered.Beta             : num  5.15 3.43 1.93 4.39 2.64 4.44 0.79 1.68 4.83 6.16 ...
##  $ Total.Levered.Beta               : num  7.36 3.76 3.09 5.07 4.9 5.24 1.88 2.37 5.71 7.18 ...
##  $ Current.PE                       : num  73 29.8 47.1 27.9 13.6 ...
##  $ Trailing.PE                      : num  30.4 31.1 28.1 27.8 15.1 ...
##  $ Forward.PE                       : num  27.5 30.9 14.4 23.9 29.6 ...
##  $ PEG.Ratio                        : num  1.71 1.65 0.31 1.43 0.59 1.11 1.69 1.38 1.4 1.89 ...
##  $ Aggregate.Mkt.Cap.Net.Income     : num  31.4 19.5 14.7 30.2 10 ...
##  $ Aggregate.Mkt.Cap.Trailing.Net.Income: num  22.4 17.8 10.9 24.3 13 ...
##  $ Number.of.Firms                  : int  52 93 22 64 22 75 13 676 22 46 ...
##  $ Average.Correlation              : num  16.1 30.8 31.6 19.6 22.3 ...
##  $ Expected.Growth.Next.5.Years     : num  13.1 10.8 34.7 17 21.9 ...
```

We first analyze the relationship between current PE, forward PE, beta, and expected growth in the next 5 years.

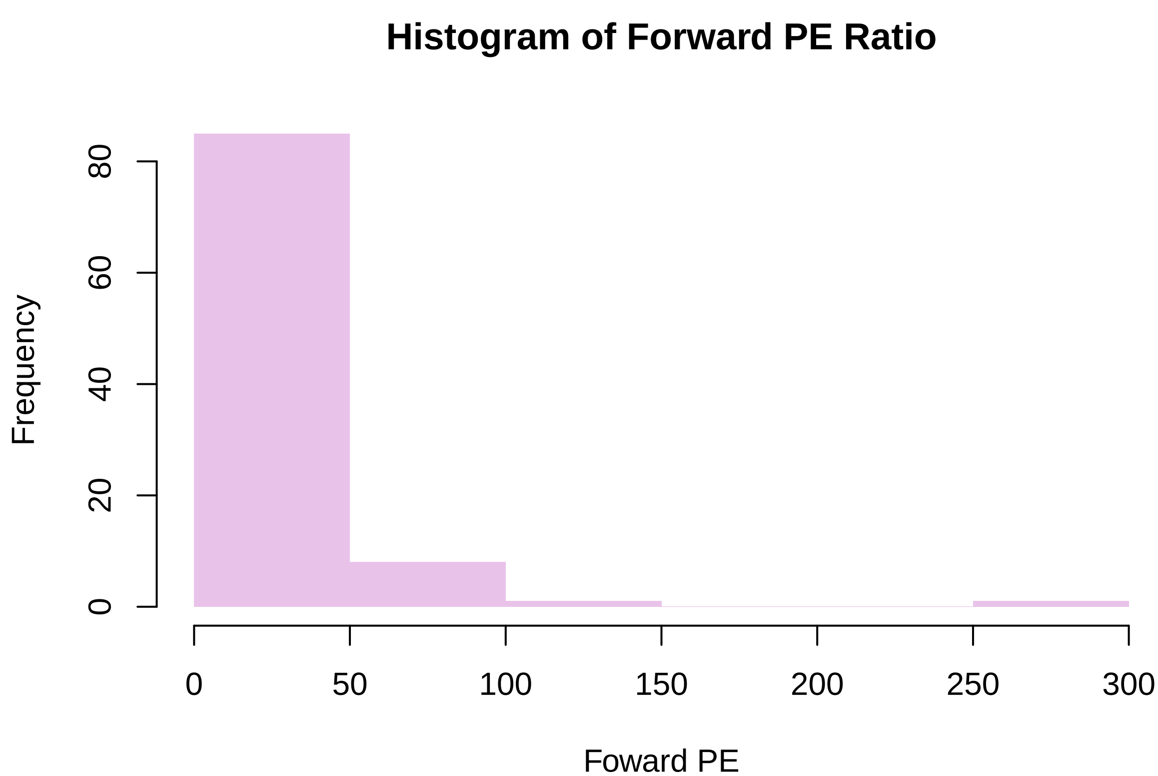
```r
# Scatter plot of Current PE to Forward PE
ggplot(clean_data, aes(x = Current.PE, y = Forward.PE)) +
    geom_point(color = rgb(100, 200, 200, 100, maxColorValue = 255)) +
    ggtitle("Current PE Ratio vs. Forward PE Ratio") +
    xlab("Current PE Ratio") + ylab("Foward PE Ratio") + theme_classic()
```



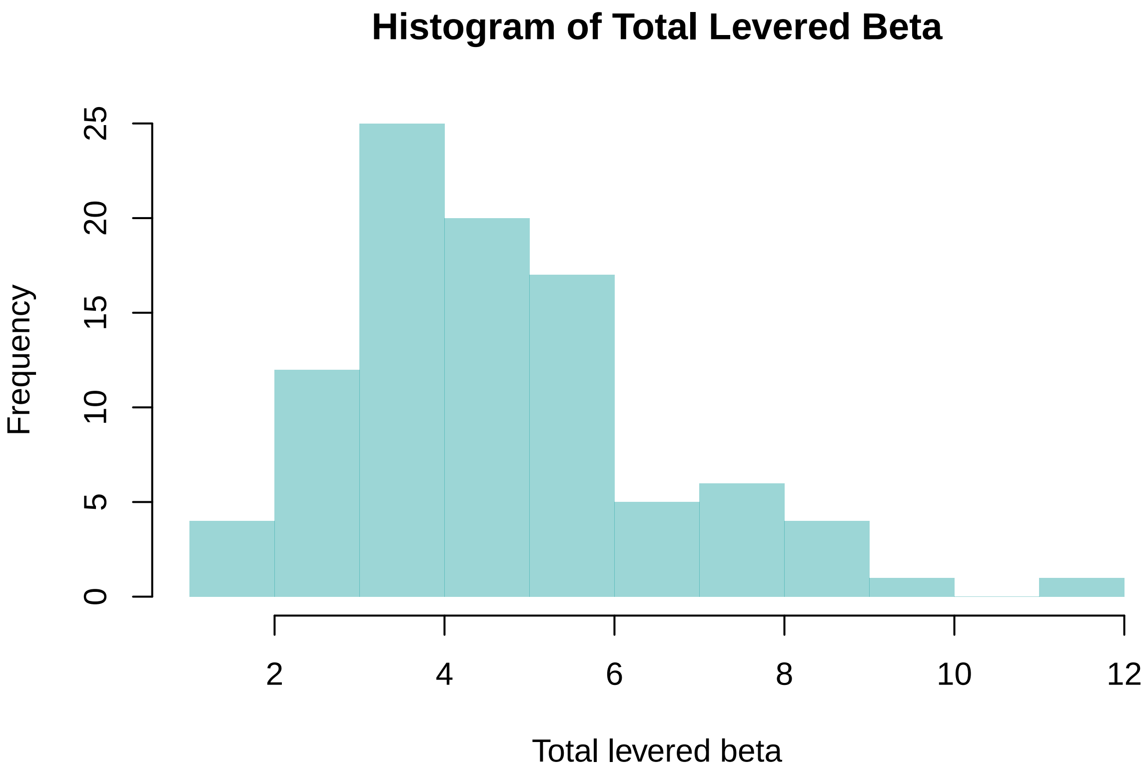

```r
# Histogram of Forward PE, Total Levered Beta, and Expected Growth in the
# Next 5 Years
hist(clean_data$Current.PE, border = NA,
     main = "Histogram of Current PE Ratio", xlab = "Current PE",
     ylab = "Frequency", col = rgb(200, 100, 100, maxColorValue = 255))
```



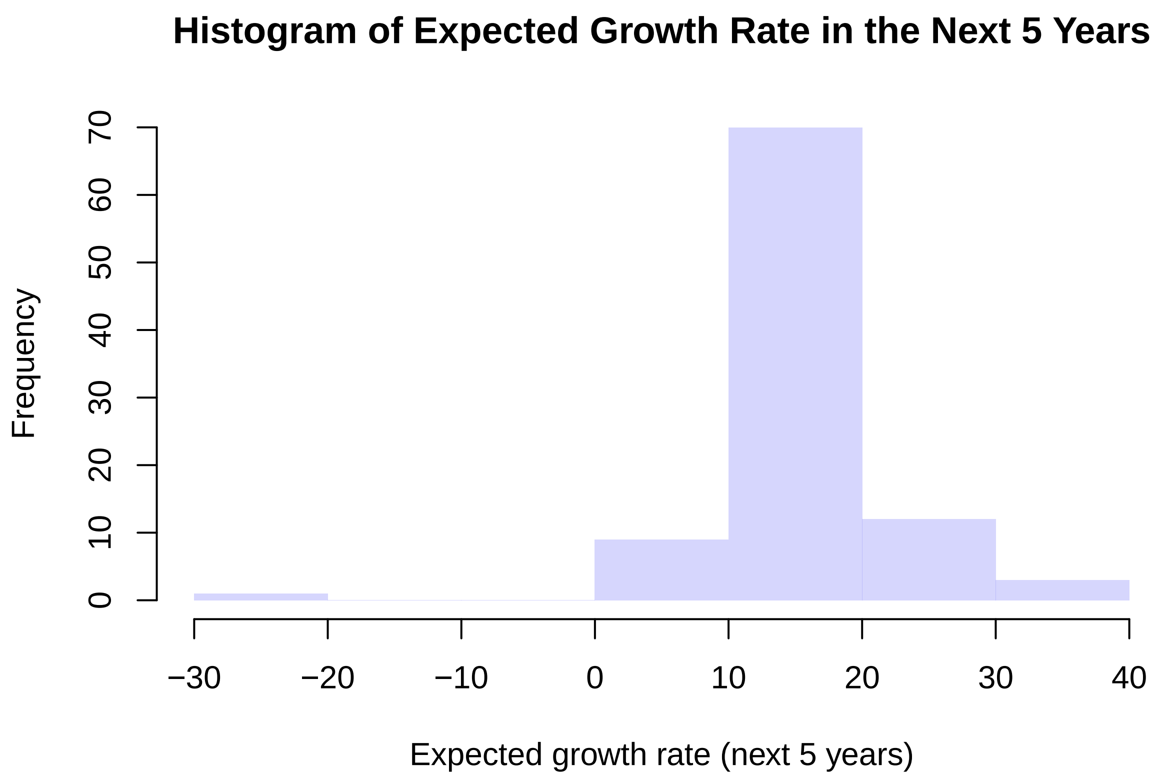

```r
hist(clean_data$Forward.PE, border = NA,
     main = "Histogram of Forward PE Ratio", xlab = "Foward PE",
     ylab = "Frequency", col = rgb(200, 100, 200, 100, maxColorValue = 255))
```



```r
hist(clean_data$Total.Levered.Beta, border = NA, xlab = "Total levered beta",
     ylab = "Frequency", col = rgb(0, 150, 150, 100, maxColorValue = 255),
     main = "Histogram of Total Levered Beta")
```




```r
hist(clean_data$Expected.Growth.Next.5.Years, border = NA,
     xlab = "Expected growth rate (next 5 years)", ylab = "Frequency",
     col = rgb(150, 150, 250, 100, maxColorValue = 255),
     main = "Histogram of Expected Growth Rate in the Next 5 Years")
```



We find that foward PE ratio is lower for each of the current PE ratios in this graph. This makes sense because the denominator of forward PE ratio is expected earnings per share while the denominator of current PE ratio is earnings per share. We know that expected earnings per share is higher than current earnings per share, so our graph makes sense.

We then wrote a function beta_interval to place our betas into 5 intervals. We find that we have 8 industries in the highest beta interval, which are companies in growing markets such as online retail, real estate, etc. Companies with the lowest beta are stable markets such as financial services or trucking.

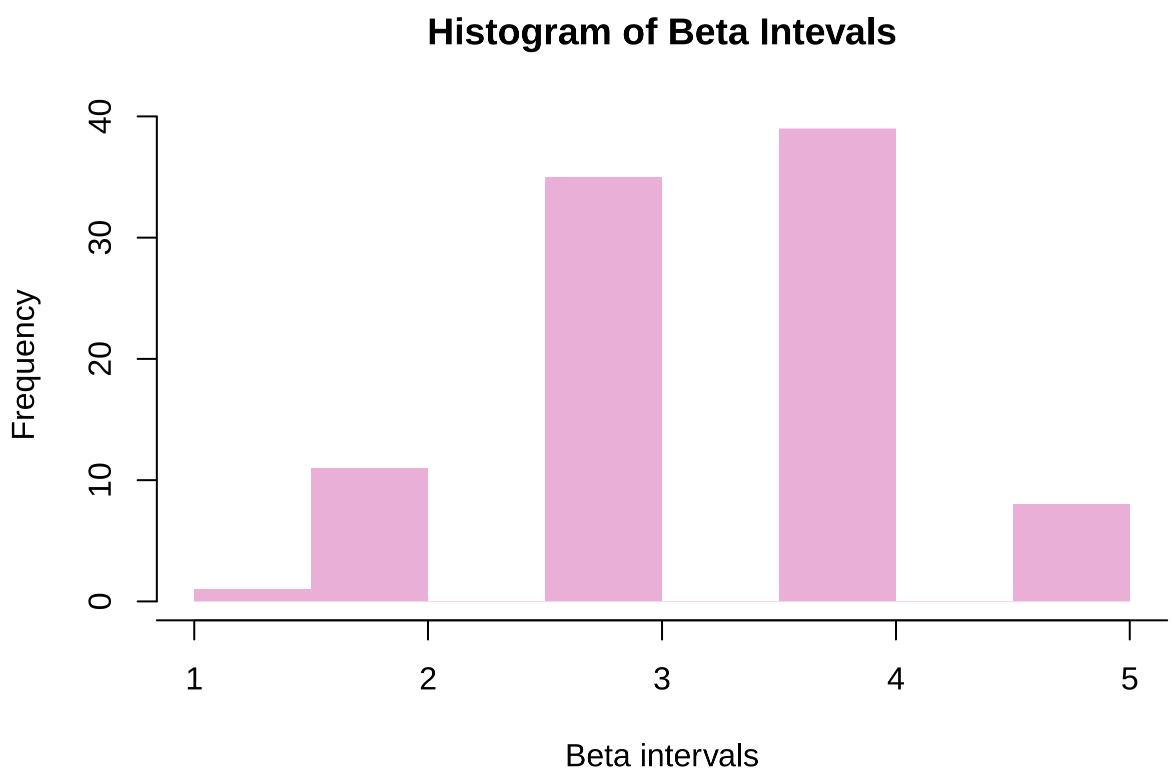
```
# Analyze which industries are in the highest and lowest beta interval
beta_intervals <- seq(from = 0, to = 1.5, by = 0.3)
interval_fun <- function(beta) {
    for (i in 1:5) {
        if (beta_intervals[i] < beta & beta < beta_intervals[i + 1]) {
            return (i)
        }
    }
}
beta_interval <- unlist(lapply(clean_data$Average.Unlevered.Beta,
                               FUN = interval_fun))
as.vector(clean_data$Industry[unlist(beta_interval) == 5])
```

```
## [1] "Construction Supplies"             "Electronics (Consumer & Office)"
## [3] "Food Wholesalers"                  "Oilfield Svcs/Equip."
## [5] "Real Estate (General/Diversified)" "Retail (Building Supply)"
## [7] "Retail (Online)"                   "Software (Internet)"
```

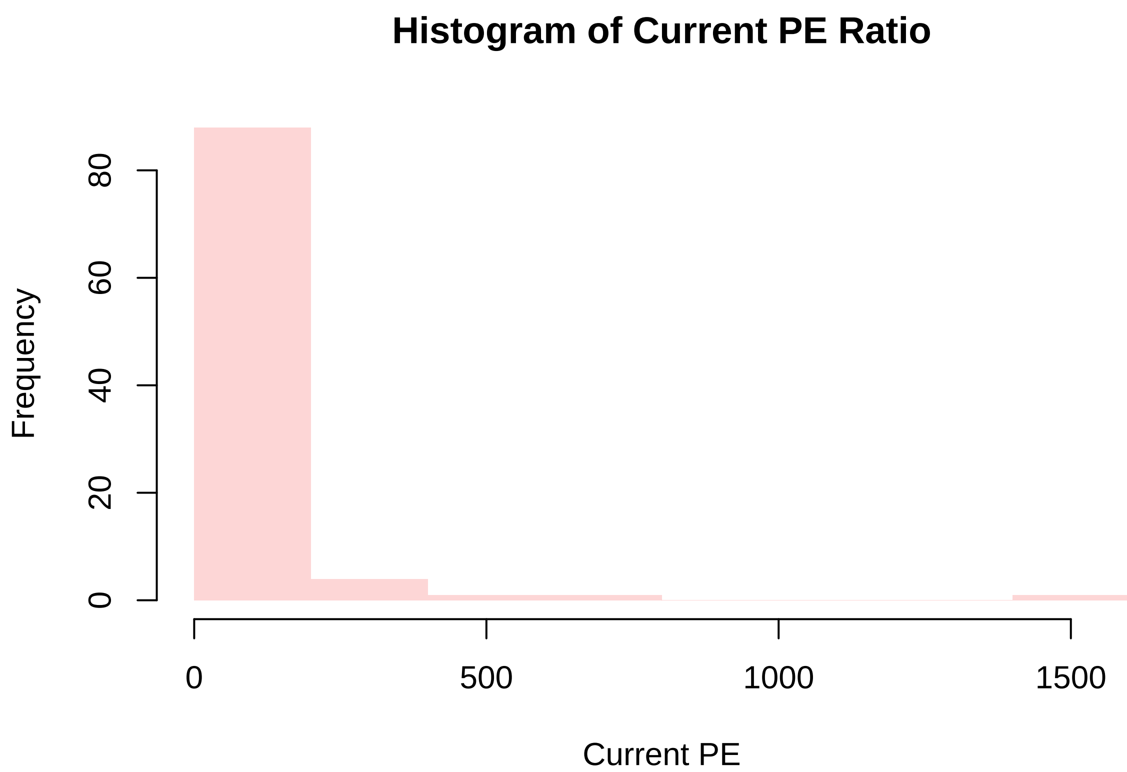
```
as.vector(clean_data$Industry[unlist(beta_interval) == 1])
```

```
## [1] "Financial Svcs. (Non-bank & Insurance)"
```

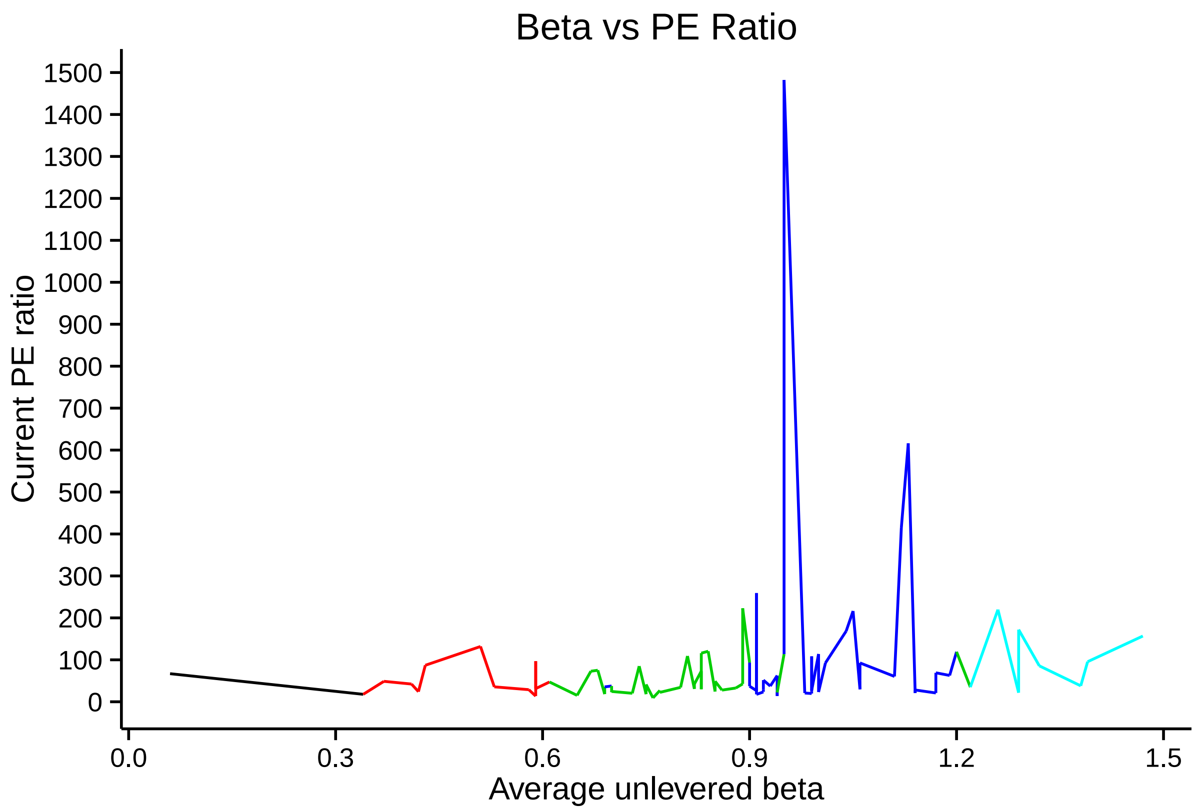
```r
# Make a histogram of the beta intervals created above
dfbeta_interval <- data.frame(table(unlist(beta_interval)))
names(dfbeta_interval) <- c("Interval", "Frequency")
hist(beta_interval, border = NA, main = "Histogram of Beta Intevals",
     xlab = "Beta intervals", ylab = "Frequency",
     col = rgb(200, 50, 150, 100, maxColorValue = 255))

# Histogram of PE Ratio and bar plot of Beta versus PE Ratio by given beta
# intervals
axis(side = 1, at = seq(0,1500,by = 300))
```




```r
hist(clean_data$Current.PE, border = NA,
     main = "Histogram of Current PE Ratio", xlab = "Current PE",
     ylab = "Frequency", col = rgb(250, 150, 150, 100, maxColorValue = 255))
```



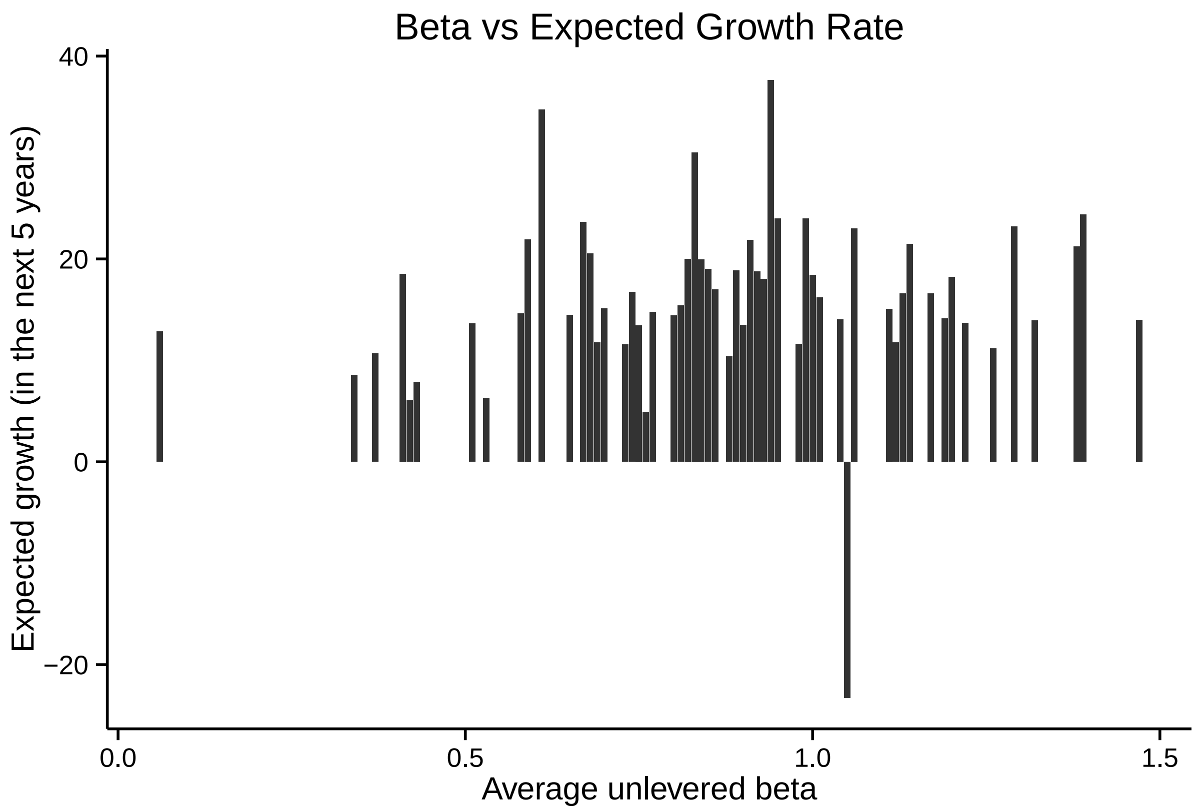

```
dat1 <- data.frame(PE = industries_only$Current.PE,
                   beta = industries_only$Average.Unlevered.Beta)
ggplot(dat1, aes(x = beta, y = PE)) + xlab("Average unlevered beta") +
    geom_line(stat = "identity", color = beta_interval) +
    ggtitle("Beta vs PE Ratio") + theme_classic() +
    scale_x_continuous(breaks = beta_intervals) + ylab("Current PE ratio") +
    scale_y_continuous(breaks = seq(from = 0, to = 1500, by = 100))
```



From the histogram, we can see that most industries have PE ratios of 0-200, with a few outliers. From the plot, we find that industries with average betas of 0.9-1.2 (Interval 4) have relatively higher PE ratios, including the outliers.

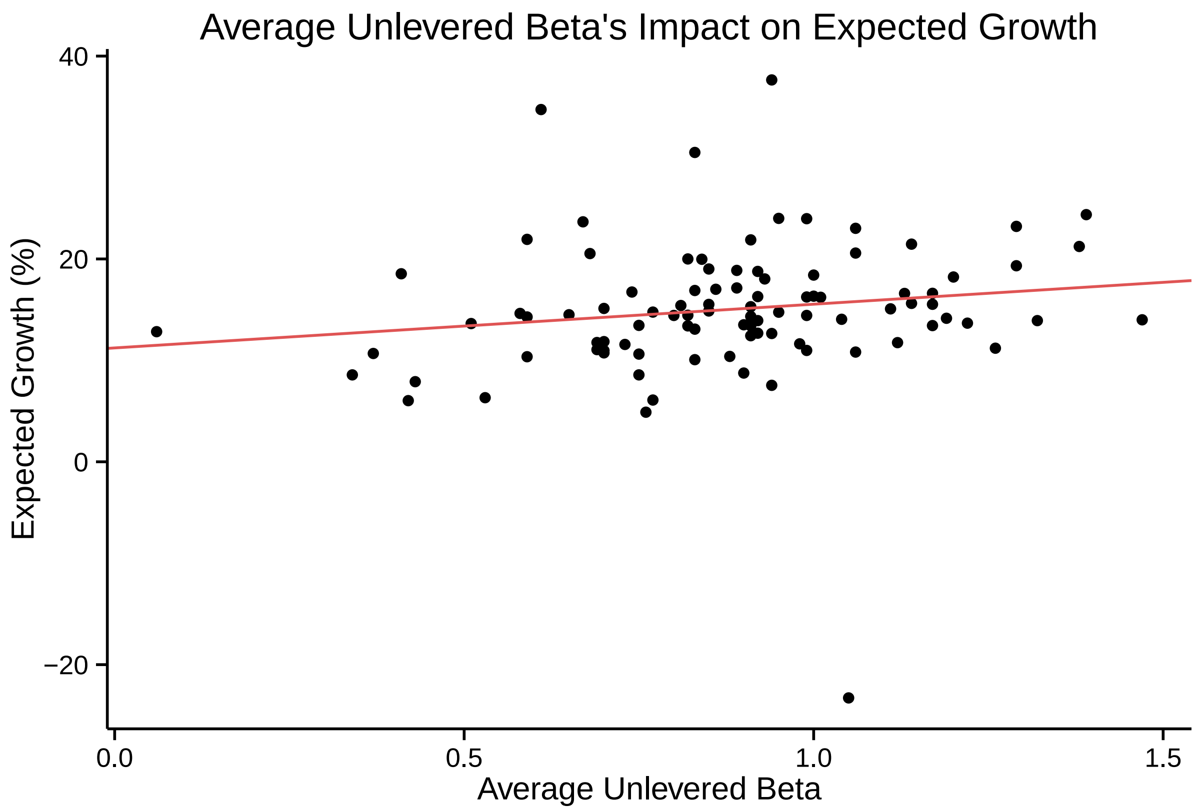
We further explored beta through making a bar plot of average unlevered beta to the expected growth rate.

```
# Bar plot of Beta versus Expected Growth Rate in the Next 5 Years
dat2 <- data.frame(beta = clean_data$Average.Unlevered.Beta,
                   growth = clean_data$Expected.Growth.Next.5.Years)
ggplot(dat2, aes(x = beta, y = growth)) + xlab("Average unlevered beta") +
    geom_bar(position = "identity", stat = "identity") +
    ggtitle("Beta vs Expected Growth Rate") + theme_classic() +
    ylab("Expected growth (in the next 5 years)")
```



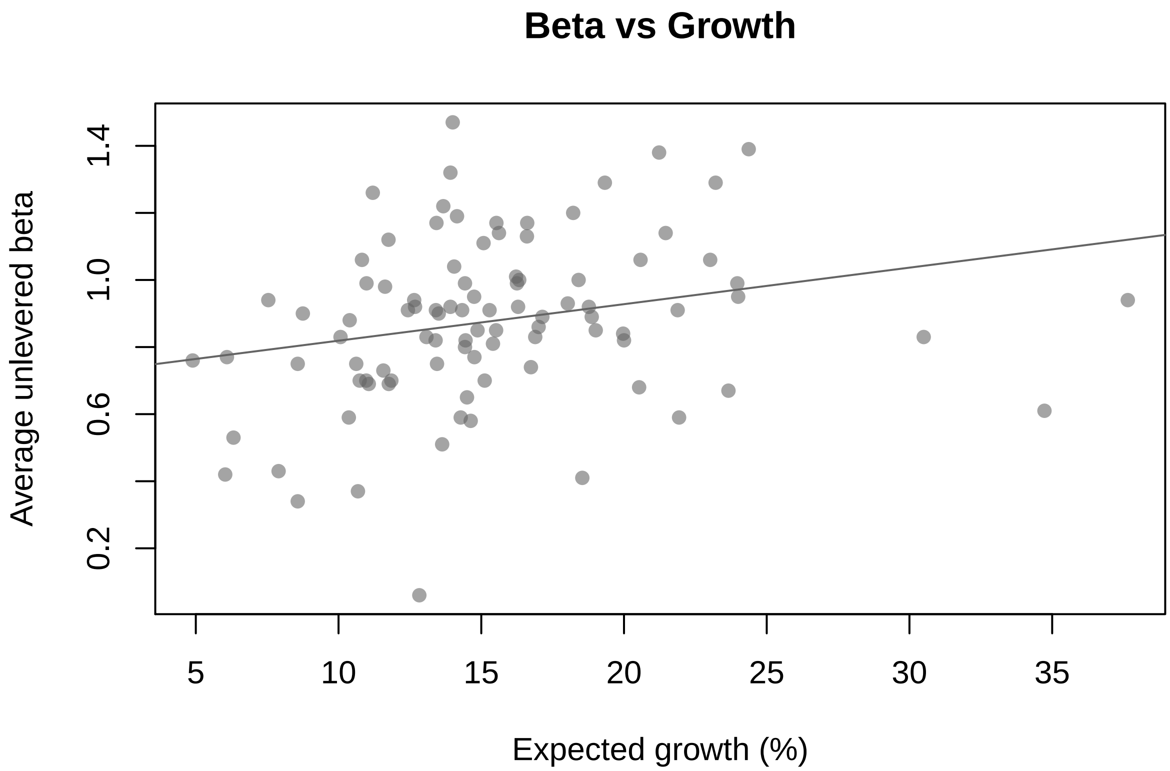
We see that industries with betas of slightly below 1 have higher expected growth. We then inspect the regression and scatterplot of beta to growth.

```
# Graph the scatterplot of unlevered beta's impact on expected growth
# when accounting for their regression
reg1 <- lm(Expected.Growth.Next.5.Years ~ Average.Unlevered.Beta,
               data = clean_data)
ggplot(clean_data, aes(x = Average.Unlevered.Beta,
                       y = Expected.Growth.Next.5.Years)) +
    geom_point() + xlab("Average Unlevered Beta") +
    ylab("Expected Growth (%)") + theme_classic() +
    geom_abline(aes(slope = unname(coef(reg1)["Average.Unlevered.Beta"]),
                    intercept = unname(coef(reg1)["(Intercept)"])),
                color = rgb(223, 84, 84, maxColorValue = 255)) +
    ggtitle("Average Unlevered Beta's Impact on Expected Growth")
```



The summary of the regression, shows that regressing unlevered beta on growth does not provide a good linear relationship between the two variables, as the coefficient does not have linear statistical significance. This is backed by the scatterplot; visually, the correlation between the two variables isn't strong. However, there are some outliers where growth is negative and beta is really low. We then create a linear regression after removing the outliers.

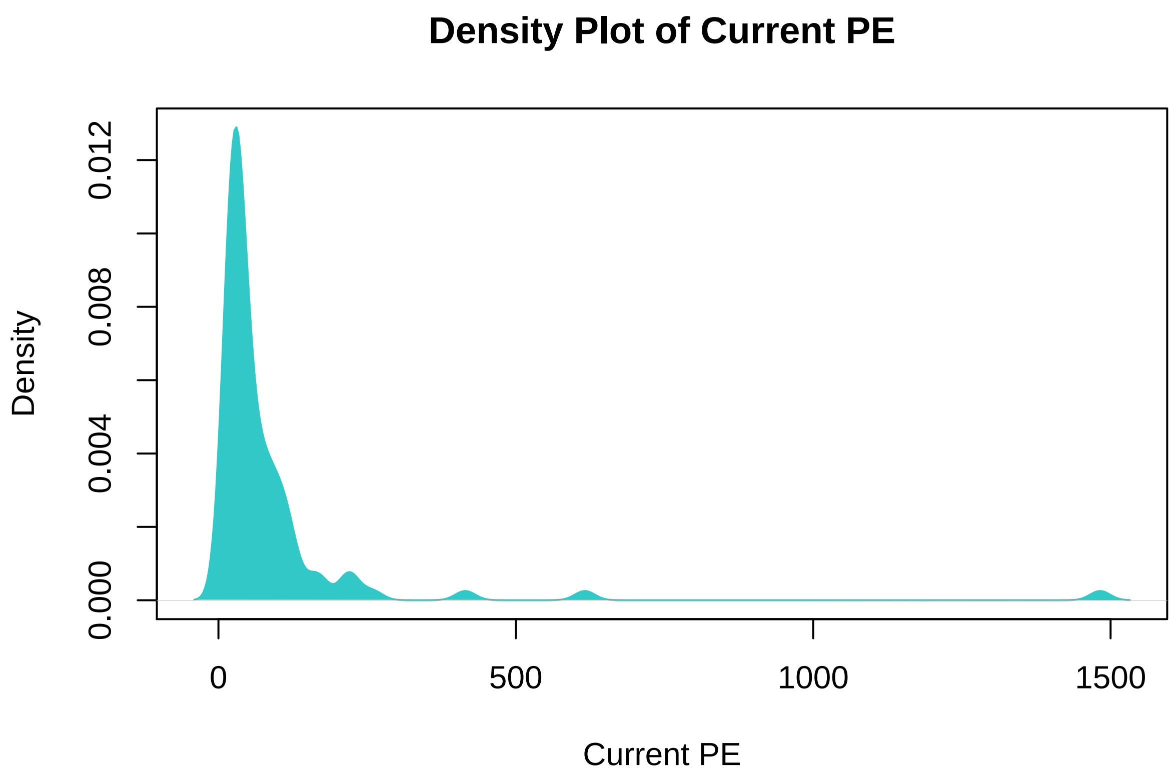
```
# Make a scatter plot and regression of beta to growth accounting
# for outliers
sorted_beta <- sort(clean_data$Average.Unlevered.Beta,
                    index.return = TRUE)[[2]]
sort_beta <- sort(clean_data$Average.Unlevered.Beta)
sorted_growth1 <- clean_data$Expected.Growth.Next.5.Years[sorted_beta]
sorted_beta_growth <- data.frame(beta = sort_beta, growth = sorted_growth1)
dat3 <- subset(sorted_beta_growth, sorted_growth1 > 0)
plot(dat3$growth, dat3$beta,
     col = rgb(100, 100, 100, 150, maxColorValue = 255), pch = 16,
     bg = rgb(200, 200, 200, maxColorValue = 255), main = "Beta vs Growth",
     xlab = "Expected growth (%)", ylab = "Average unlevered beta")
fit_growth_beta <- lm(beta ~ growth, data = dat3)
abline(fit_growth_beta, col = rgb(100, 100, 100, maxColorValue = 255))
```



There is positive correlation between beta and growth. This makes sense because growing industries are probably taking more risks and therefore are more sensitive to changes in the market, whereas mature industries would not be as sensitive to market fluctuations.

We also further explore the relationship between current PE and expected growth through multiple methods.

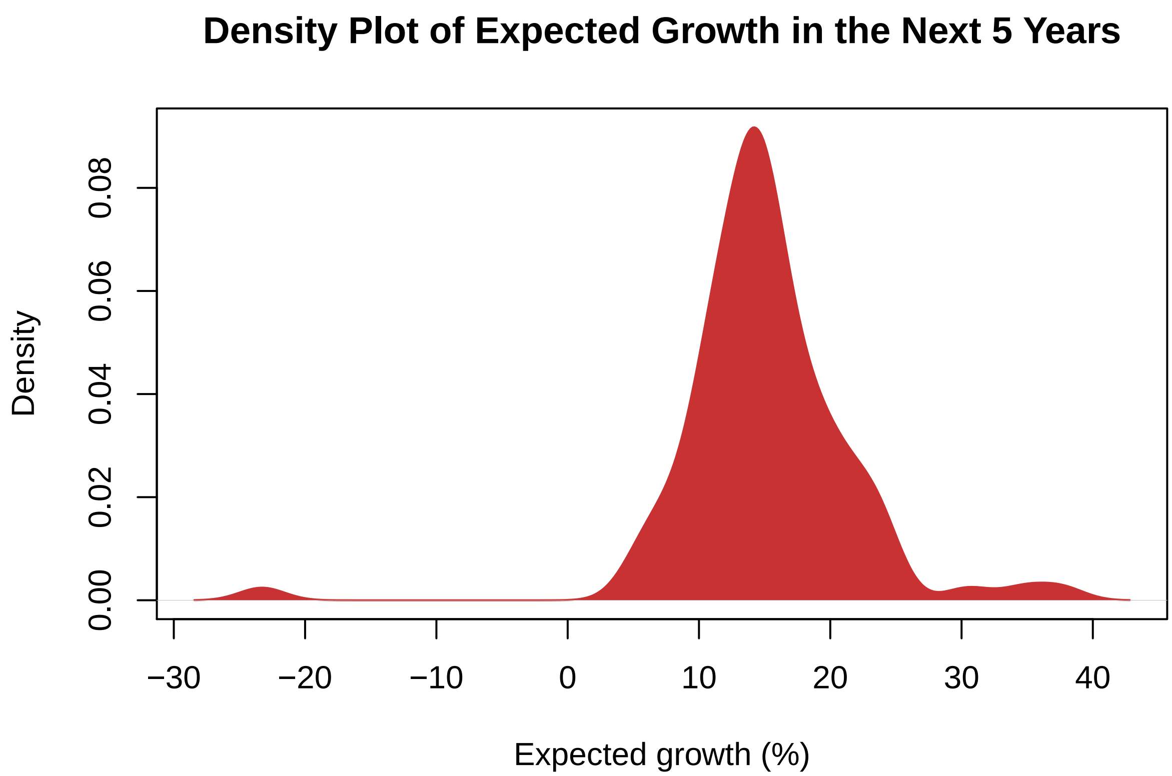
```r
# Density curve of Current PE and of Expected Growth in the Next 5 Years
plot(density(clean_data$Current.PE), main = "Density Plot of Current PE",
     xlab = "Current PE", ylab = "Density",
     col = rgb(50, 200, 200, maxColorValue = 255))
polygon(density(clean_data$Current.PE), border = NA,
        col = rgb(50, 200, 200, maxColorValue = 255))
```



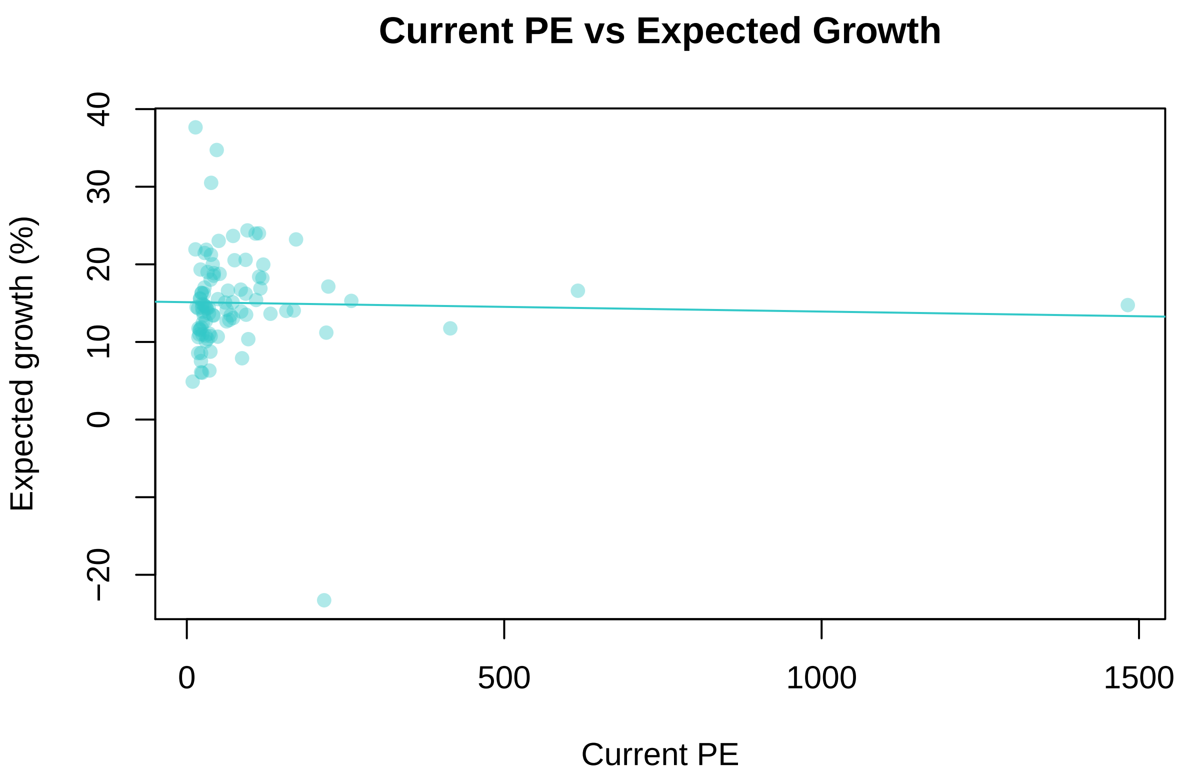

```
plot(density(clean_data$Expected.Growth.Next.5.Years),
     main = "Density Plot of Expected Growth in the Next 5 Years",
     xlab = "Expected growth (%)", ylab = "Density", border = NA,
     col = rgb(200, 50, 50, maxColorValue = 255))
```

```
## Warning in plot.window(...): "border" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "border" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "border" is
## not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "border" is
## not a graphical parameter

## Warning in box(...): "border" is not a graphical parameter

## Warning in title(...): "border" is not a graphical parameter
```

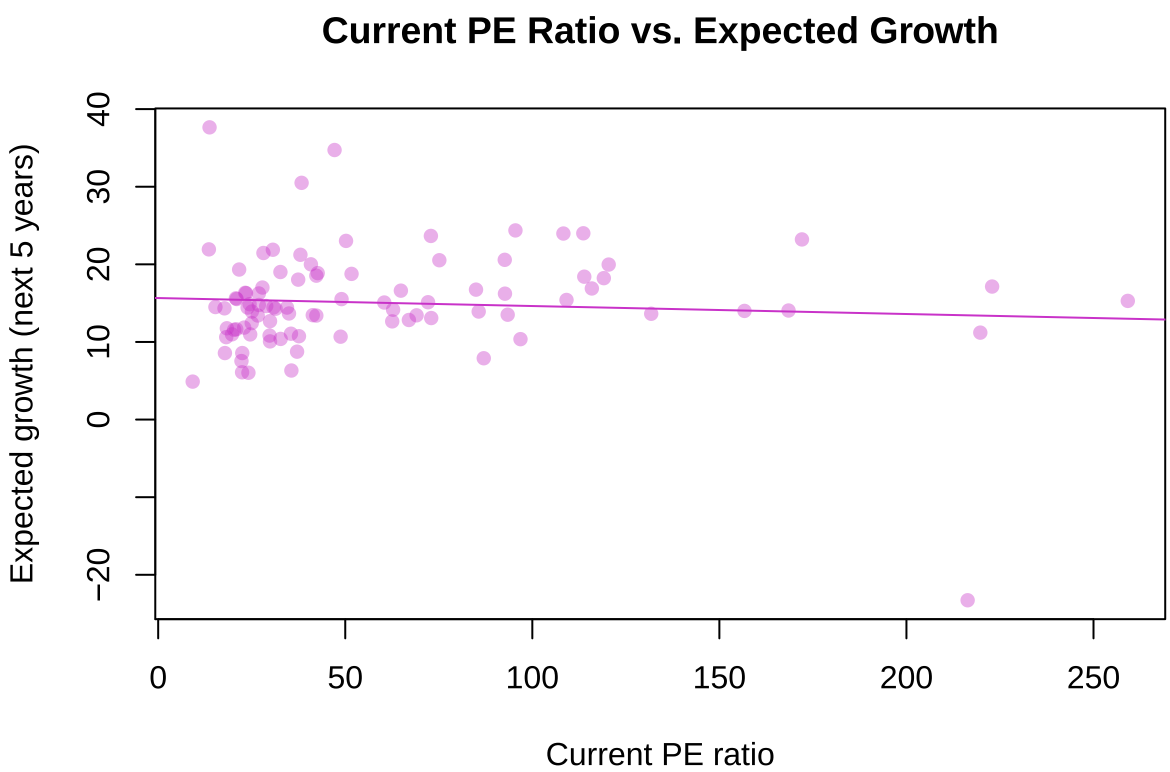
```
polygon(density(clean_data$Expected.Growth.Next.5.Years),
        border = NA, col = rgb(200, 50, 50, maxColorValue = 255))
```




```r
# Scatter plot of Current PE to Expected Growth in the Next 5 Years
plot(clean_data$Current.PE, clean_data$Expected.Growth.Next.5.Years,
     col = rgb(50, 200, 200, 100, maxColorValue = 255), pch = 16,
     bg = rgb(200, 200, 200, maxColorValue = 255), xlab = "Current PE",
     ylab = "Expected growth (%)", main = "Current PE vs Expected Growth")

# Make a linear regression of Current PE to Expected Growth in the Next 5 Years
fit1 <- lm(Expected.Growth.Next.5.Years ~ Current.PE, data = clean_data)
abline(fit1, col = rgb(50, 200, 200, maxColorValue = 255))
```

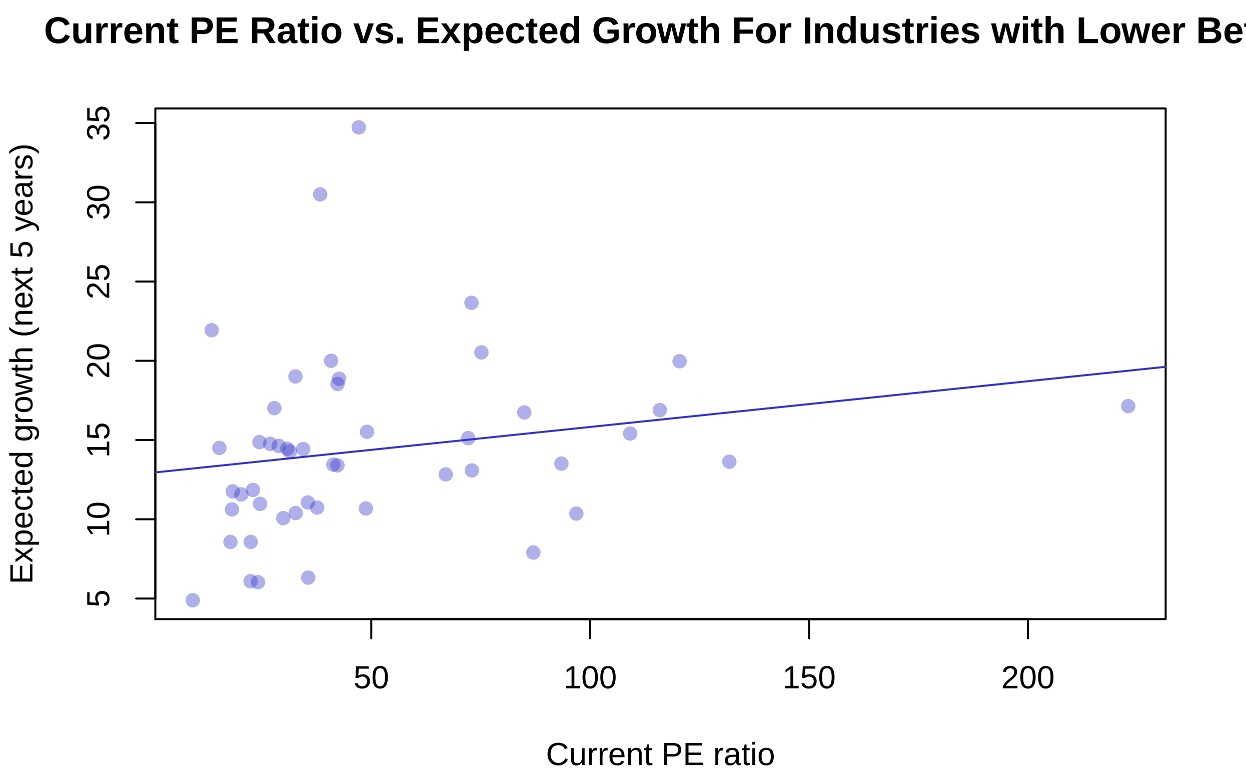



```
# Remove outliers and find the linear regression of Current PE to Expected
# Growth in the Next 5 Years
sorted_pe <- sort(clean_data$Current.PE, decreasing = TRUE,
                  index.return = TRUE)[[2]]
sort_pe <- sort(clean_data$Current.PE, decreasing = TRUE)
sorted_growth2 <- clean_data$Expected.Growth.Next.5.Years[sorted_pe]
sorted_pe_growth <- data.frame(pe = sort_pe, growth = sorted_growth2)
dat4 <- subset(sorted_pe_growth, sort_pe < 300)
plot(dat4$pe, dat4$growth, main = "Current PE Ratio vs. Expected Growth",
     xlab = "Current PE ratio", ylab = "Expected growth (next 5 years)",
     col = rgb(200, 50, 200, 100, maxColorValue = 255), pch = 16)
fit2 <- lm(growth ~ pe, data = dat4)
abline(fit2, col = rgb(200, 50, 200, maxColorValue = 255))
```

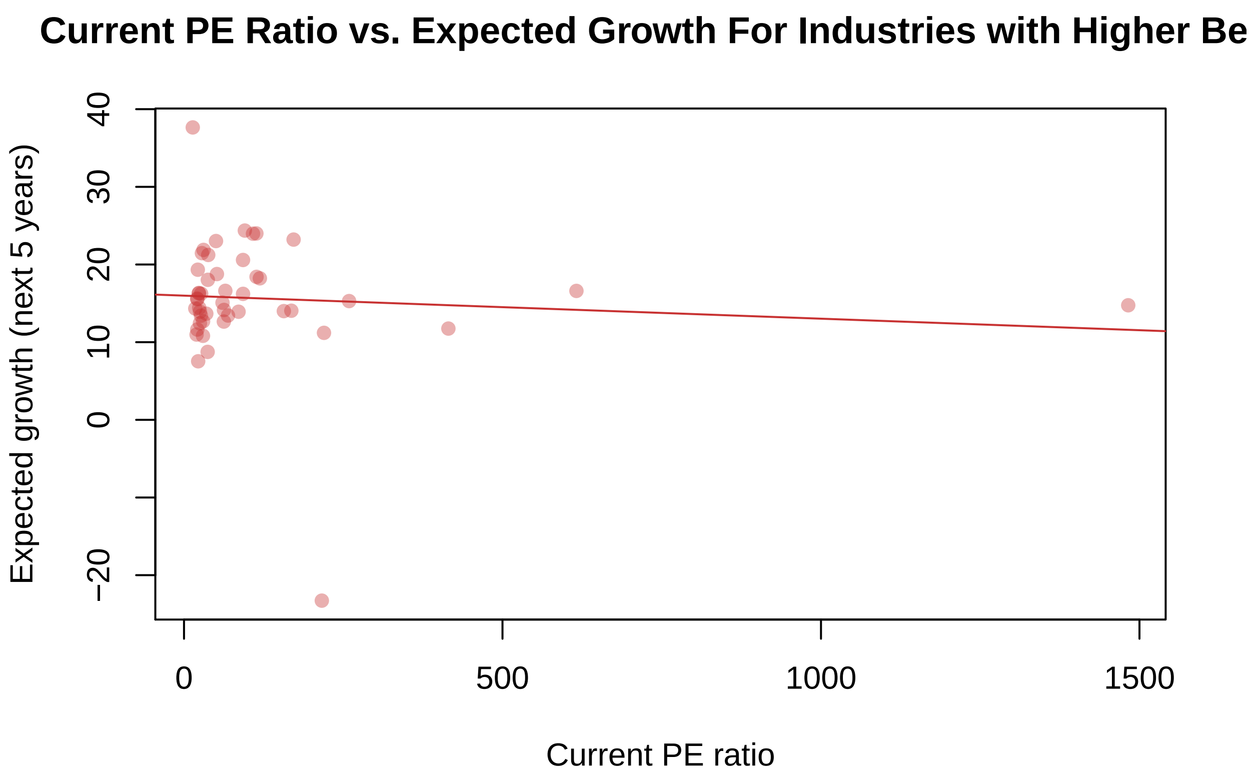


Simple methods, such as density curves, scatterplots, and regressions reveal that there is not much initial correlation between current PE and expected growth, so we account for average unlevered beta.

```
# Account for beta when sorting PE ratios and growth by taking the PE ratios
# and growths of industries with lower beta
beta_pe <- clean_data$Current.PE[sorted_beta]
beta_growth <- clean_data$Expected.Growth.Next.5.Years[sorted_beta]
end <- length(beta_pe)
cutoff <- round(end/2)
beta1 <- data.frame(pe = beta_pe[1:cutoff], growth = beta_growth[1:cutoff])
plot(beta1$pe, beta1$growth,
     main = paste("Current PE Ratio vs. Expected Growth",
                  "For Industries with Lower Beta"),
     xlab = "Current PE ratio", ylab = "Expected growth (next 5 years)",
     col = rgb(50, 50, 200, 100, maxColorValue = 255), pch = 16)
fit3 <- lm(growth ~ pe, data = beta1)
abline(fit3, col = rgb(50, 50, 200, maxColorValue = 255))
```



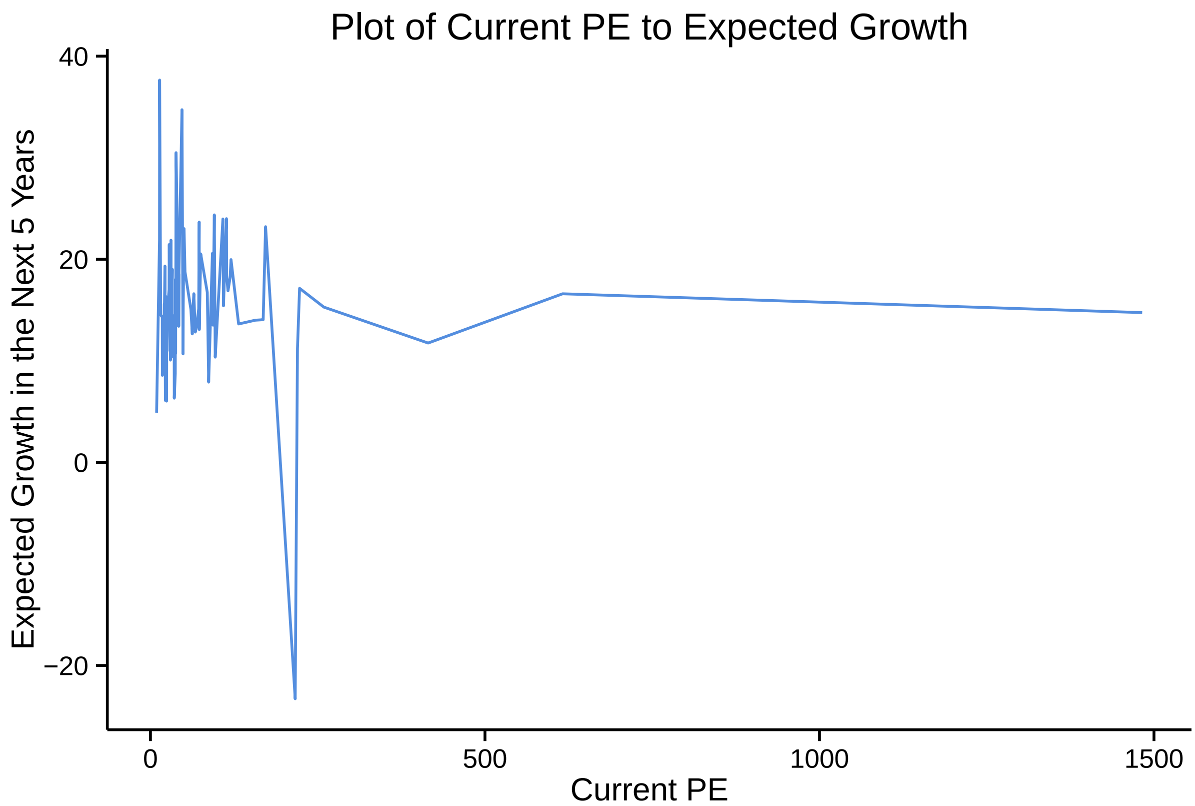

```r
# Account for beta when sorting PE ratios and growth by taking the PE ratios
# and growths of industries with higher beta
beta2 <- data.frame(pe = beta_pe[(cutoff+1):end],
                    growth = beta_growth[(cutoff+1):end])
plot(beta2$pe, beta2$growth,
     main = paste("Current PE Ratio vs. Expected Growth",
                                     "For Industries with Higher Beta"),
     xlab = "Current PE ratio", ylab = "Expected growth (next 5 years)",
     col = rgb(200, 50, 50, 100, maxColorValue = 255), pch = 16)
fit4 <- lm(growth ~ pe, data = beta2)
abline(fit4, col = rgb(200, 50, 50, maxColorValue = 255))
```



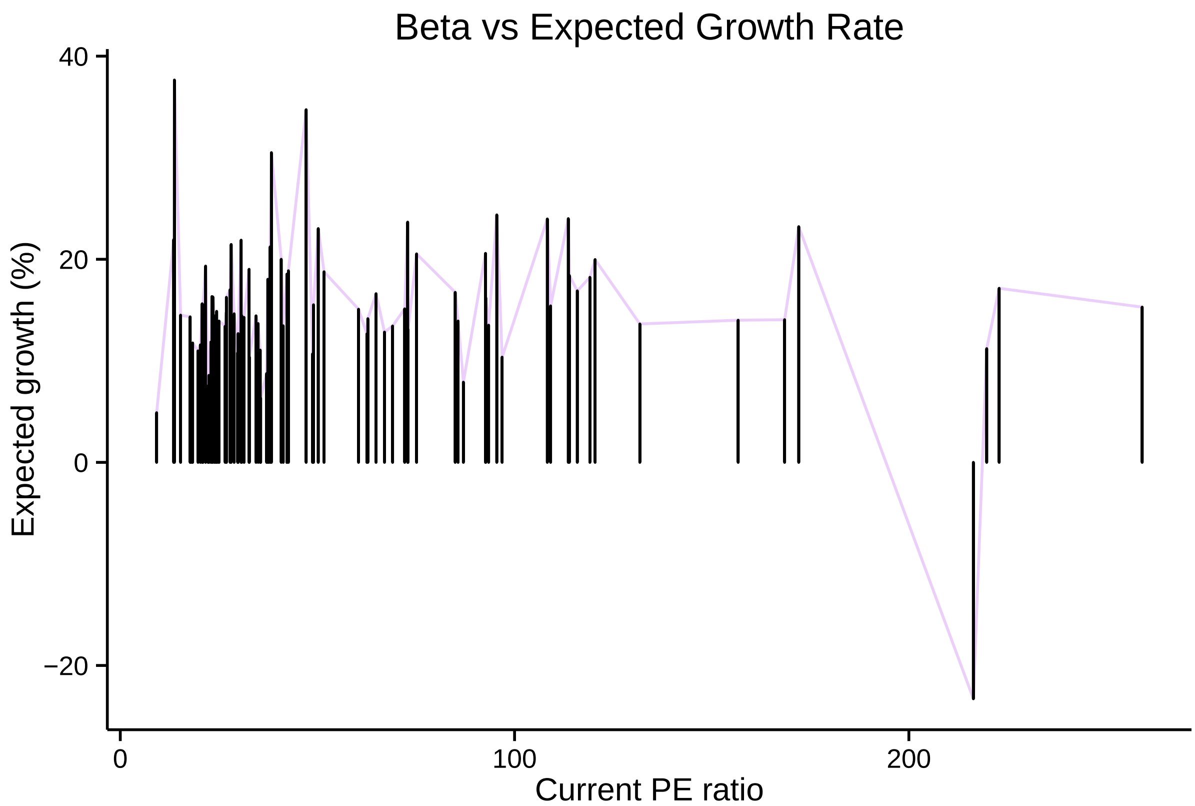
We see that industries with lower beta (risk) have a positive correlation between current PE and expected growth, while industries with higher beta have a slight negative correlation between current PE and growth. This makes sense, because established companies that have higher PE ratios are most likely those that are successful and will continue to grow in the next 5 years while companies that are just starting up will vary in terms of growth regardless of their current PE ratio.

To extract more information from current PE to growth, we made line graphs and bar plots.

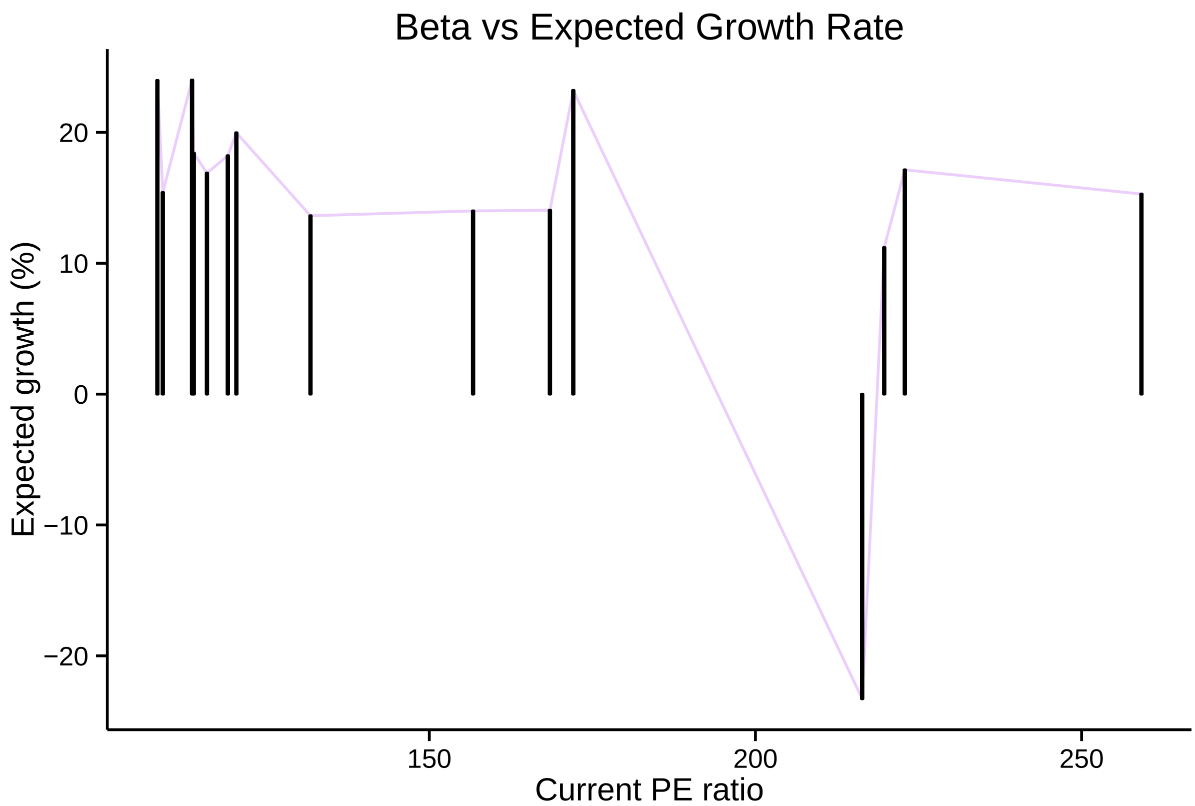
```
# Graph Current PE to Expected Growth in the Next 5 Years
ggplot(data = clean_data, aes(x = Current.PE,
                              y = Expected.Growth.Next.5.Years)) +
    geom_line(color = "#548edf") + xlab("Current PE") +
    ylab("Expected Growth in the Next 5 Years") +
    ggtitle("Plot of Current PE to Expected Growth") + theme_classic()
```



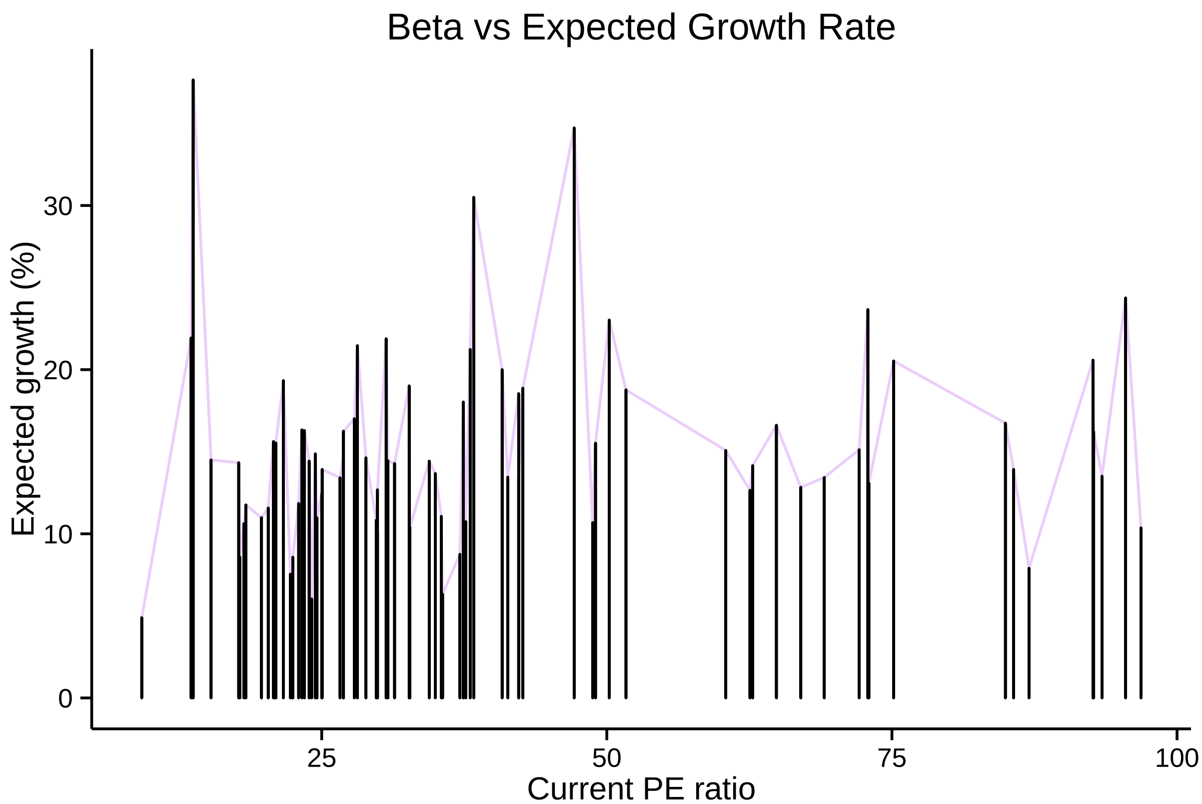

```r
# Make a bar plot of Current PE to Expected Growth for industries with
# pe, p, fulfilling p < 300
ggplot(dat4, aes(x = pe, y = growth)) + geom_line(color = "#ebcefb") +
    geom_bar(position = "identity", stat = "identity", color = "#000000") +
    ggtitle("Beta vs Expected Growth Rate") + xlab("Current PE ratio") +
    ylab("Expected growth (%)") + theme_classic()
```




```
# Make a bar plot of Current PE to Expected Growth for industries with
# pe, p, fulfilling 100 <= p < 200
dat5 <- subset(sorted_pe_growth, pe >= 100 & pe < 300)
ggplot(dat5, aes(x = pe, y = growth)) + geom_line(color = "#ebcefb") +
    geom_bar(position = "identity", stat = "identity", color = "#000000") +
    ggtitle("Beta vs Expected Growth Rate") + xlab("Current PE ratio") +
    ylab("Expected growth (%)") + theme_classic()
```

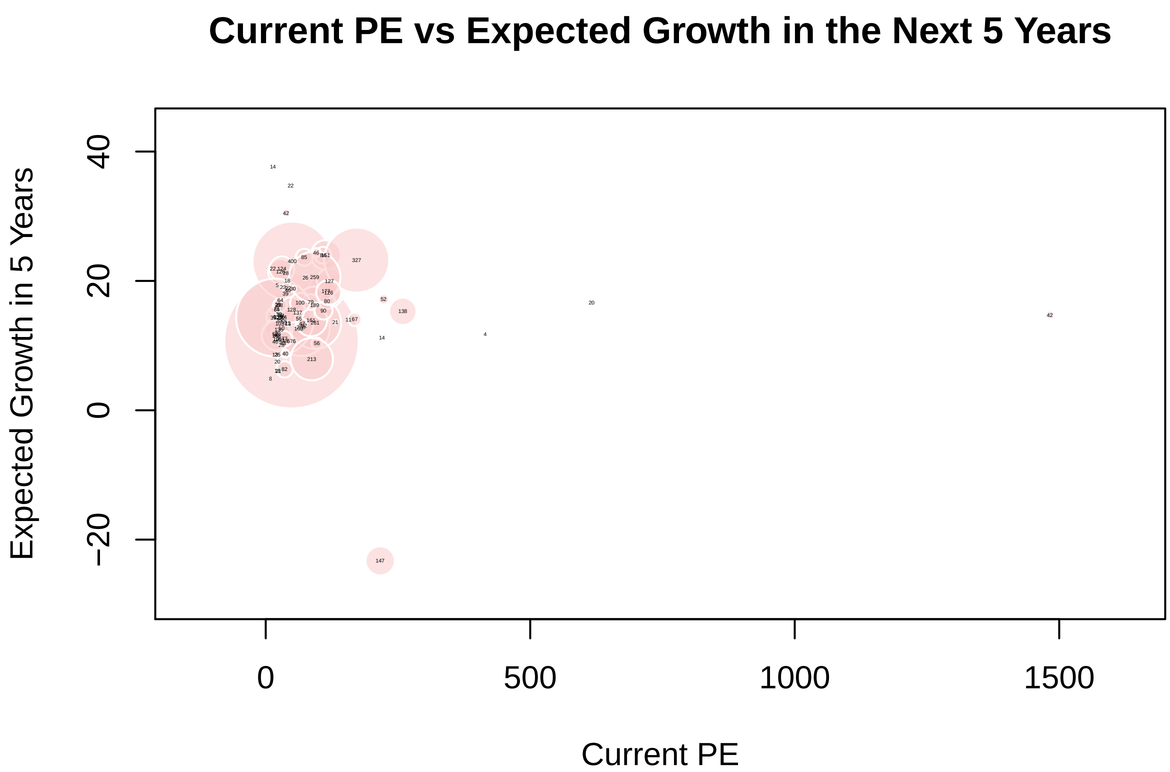



```r
# Make a bar plot of Current PE to Expected Growth for industries with
# pe, p, fulfilling p < 100
dat6 <- subset(sorted_pe_growth, pe < 100)
ggplot(dat6, aes(x = pe, y = growth)) + geom_line(color = "#ebcefb") +
    geom_bar(position = "identity", stat = "identity",
             color = "#000000") +
    ggtitle("Beta vs Expected Growth Rate") + xlab("Current PE ratio") +
    ylab("Expected growth (%)") + theme_classic()
```

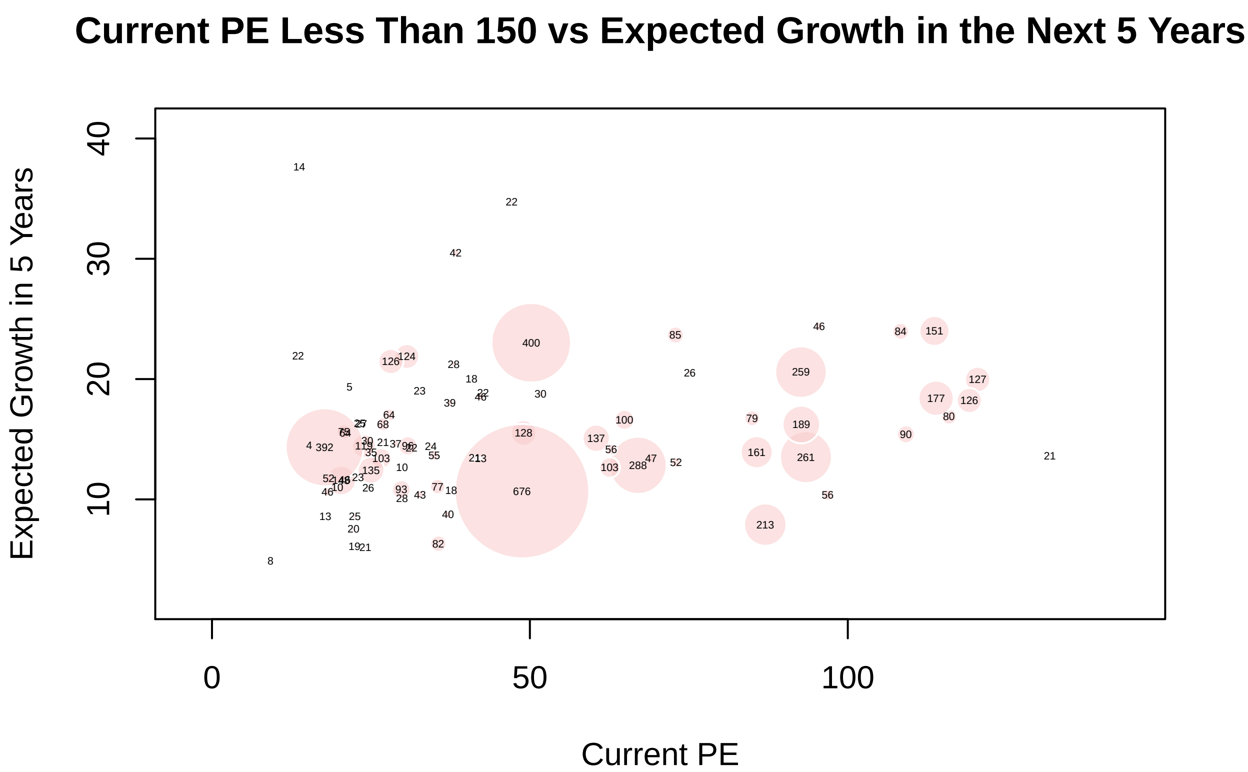


We find that most industries have current PEs centered around 0 to 50, and that those with the highest growth tend to have PEs of 50 to 100. We finally make a bubble plot of current PE to expected growth, with circles corresponding to the size of the industry.

```
# Make bubble plots of Current PE and Expected Growth in the
# Next 5 Years for different industries
symbols(industries_only$Current.PE,
        industries_only$Expected.Growth.Next.5.Years,
        circles = industries_only$Number.of.Firms,
        inches = 0.35, fg="white", bg="#faccc8d",
        xlab="Current PE", ylab="Expected Growth in 5 Years",
        main = "Current PE vs Expected Growth in the Next 5 Years")
text(industries_only$Current.PE, industries_only$Expected.Growth.Next.5.Years,
     labels = industries_only$Number.of.Firms, cex = 0.2)
```



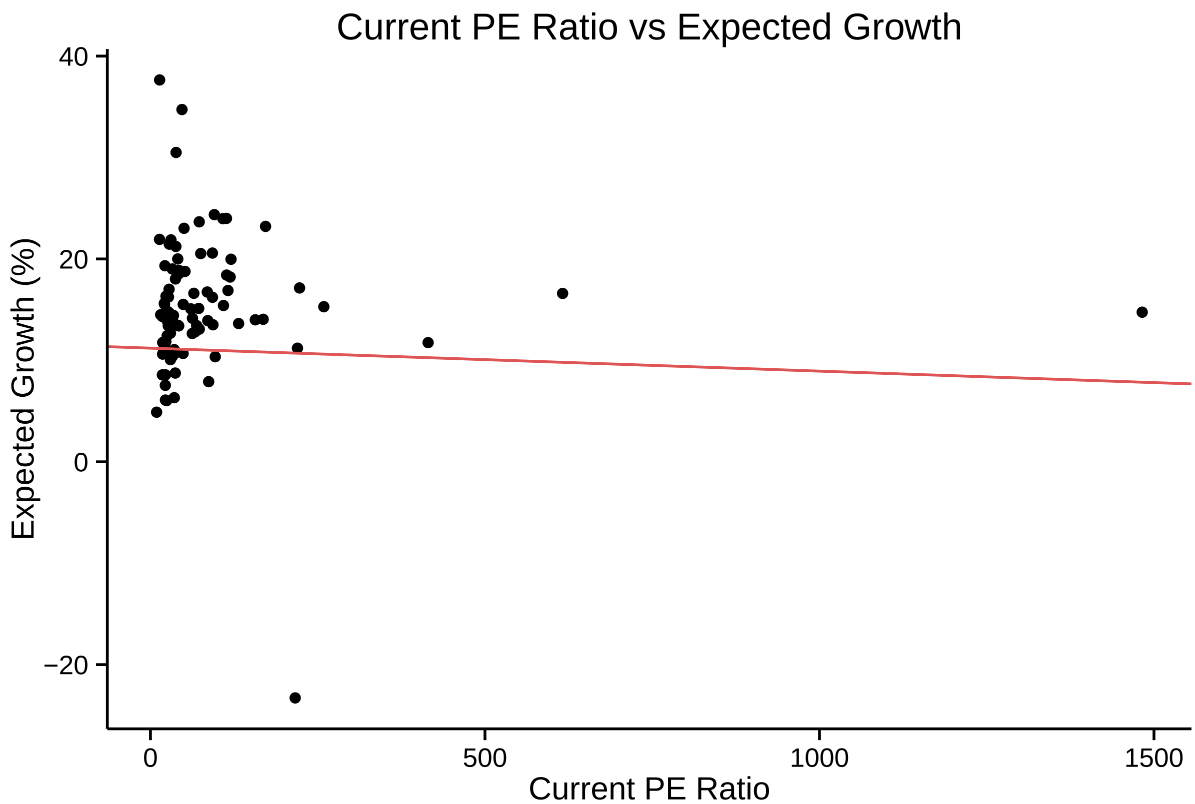

```r
# Make bubble plots of Current PE and Expected Growth in the
# Next 5 Years for different industries accounting for outliers
industry_pe <- unlist(sort(industries_only$Current.PE, decreasing = TRUE))
industry_indices <- unlist(sort(industries_only$Current.PE, decreasing = TRUE,
                        index.return = TRUE)[[2]])
industry_growth <- industries_only$
    Expected.Growth.Next.5.Years[industry_indices]
industry_num_firms <- industries_only$Number.of.Firms[industry_indices]
industry_pe_growth <- data.frame(pe = industry_pe, growth = industry_growth,
                                 num_firms = industry_num_firms)
dat7 <- subset(industry_pe_growth, industry_pe < 150)
symbols(dat7$pe, dat7$growth, circles = dat7$num_firms,
        inches=0.35, fg="white", bg="#faccc8d",
        xlab="Current PE", ylab="Expected Growth in 5 Years",
        main = paste("Current PE Less Than 150 vs",
                     "Expected Growth in the Next 5 Years"))
text(dat7$pe, dat7$growth, labels = dat7$num_firms, cex = 0.3)
```



We see that there are several outliers for current PE, but when we discount for those, we find that larger industries have PEs of 50 to 100 while smaller industries tend to have smaller values.

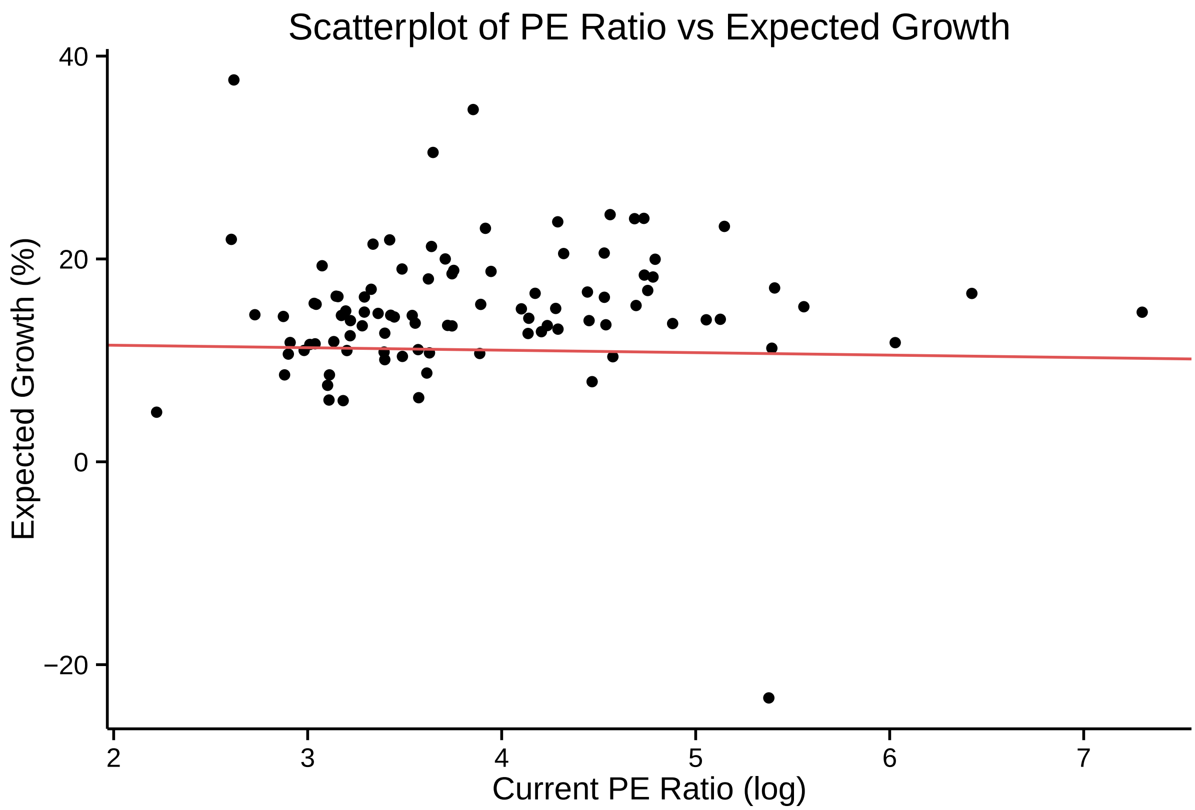
Further relationships were investigated to control for other variables that may have caused a deviation in the coefficient for the regression of unlevered beta and expected growth. One of these variables is the PE ratio.

```
# Find the multivariate regression of expected growth to unlevered beta
# and current PE ratio
reg2 <- lm(Expected.Growth.Next.5.Years ~ Average.Unlevered.Beta +
               Current.PE, data = clean_data)

# Graph the scatterplot of current PE ratio versus expected growth
ggplot(clean_data, aes(x = Current.PE, y = Expected.Growth.Next.5.Years)) +
    geom_point() + xlab("Current PE Ratio") + ylab("Expected Growth (%)") +
    geom_abline(aes(slope = unname(coef(reg2)["Current.PE"]),
                    intercept = unname(coef(reg2)["(Intercept)"])),
                color = rgb(223, 84, 84, maxColorValue = 255)) +
    ggtitle("Current PE Ratio vs Expected Growth") + theme_classic()
```

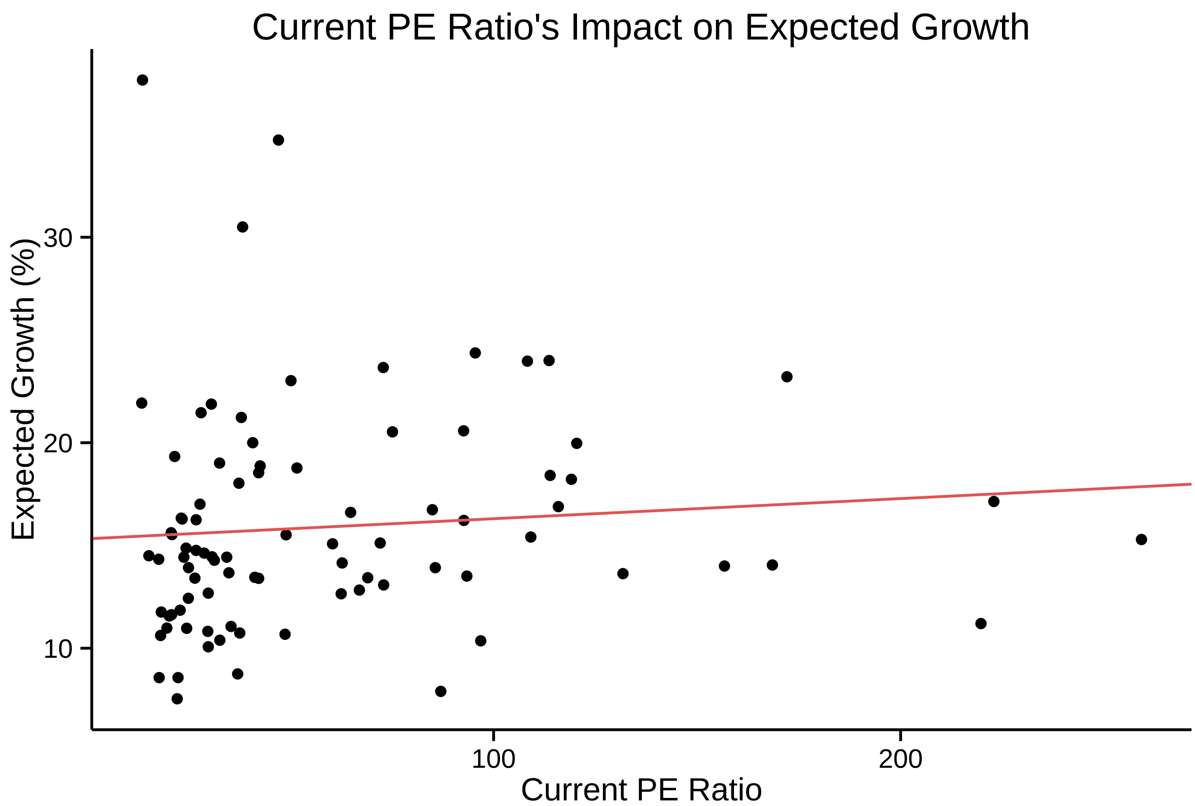


Again, this regression demonstrates that there is not a significant linear relationship between these variables, as the only statistically significant coefficient is that of the intercept. Taking the log of the independent variable yields a better looking graph with the absence of an outlier. We also manually remove the outlier.

```r
# Graph the scatterplot of the log of the current PE ratio to the growth rate
# with the corresponding multivariate regression
log_reg2 <- lm(Expected.Growth.Next.5.Years ~ log(Current.PE) +
                   Average.Unlevered.Beta, data = clean_data)
ggplot(clean_data, aes(x = log(Current.PE),
                       y = Expected.Growth.Next.5.Years)) +
    geom_point() + ggtitle("Scatterplot of PE Ratio vs Expected Growth") +
    xlab("Current PE Ratio (log)") + ylab("Expected Growth (%)") +
    geom_abline(aes(slope = unname(coef(log_reg2)["log(Current.PE)"]),
                    intercept = unname(coef(log_reg2)["(Intercept)"])),
                color = rgb(223, 84, 84, maxColorValue = 255)) +
    theme_classic()
```



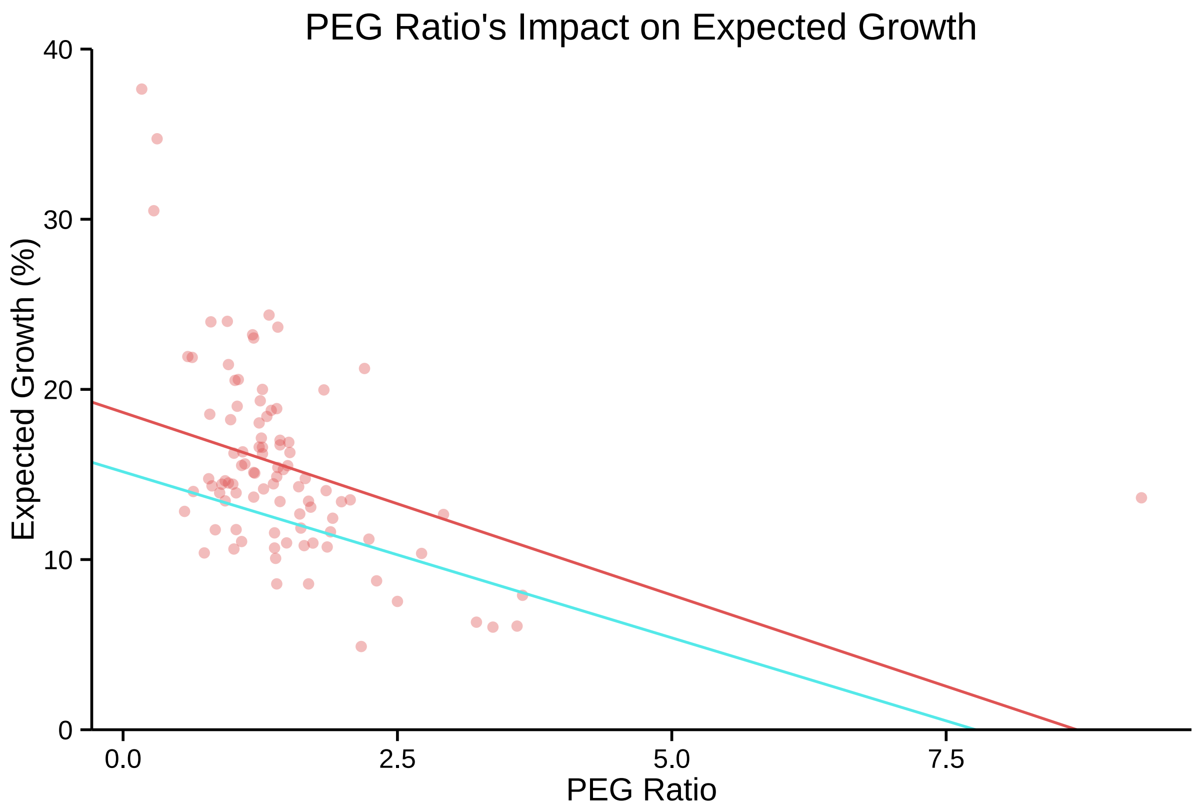

```r
# Remove the outliers of expected growth and current PE from the graph
indices_no_outliers <- sort(clean_data$Expected.Growth.Next.5.Years,
                            index.return = TRUE)[[2]]
growth_no_outliers <- unlist(sort(clean_data$Expected.Growth.Next.5.Years))
pe_no_outliers <- clean_data$Current.PE[indices_no_outliers]
dat8 <- data.frame(growth = growth_no_outliers, pe = pe_no_outliers)
dat8 <- subset(dat8, growth_no_outliers > 6.32 & pe_no_outliers < 300)

# Make a multivariate regression of growth to PE when there are no outliers
reg2_no_outliers <- lm(growth ~ pe , data = dat8)

# Plot the scatterplot and regression of current PE to expected growth
ggplot(dat8, aes(x = pe, y = growth)) + theme_classic() +
    geom_point() + xlab("Current PE Ratio") + ylab("Expected Growth (%)") +
    geom_abline(aes(slope = unname(coef(reg2_no_outliers)["pe"]),
                    intercept = unname(coef(reg2_no_outliers)["(Intercept)"])),
                color = rgb(223, 84, 84, maxColorValue = 255)) +
    ggtitle("Current PE Ratio's Impact on Expected Growth")
```



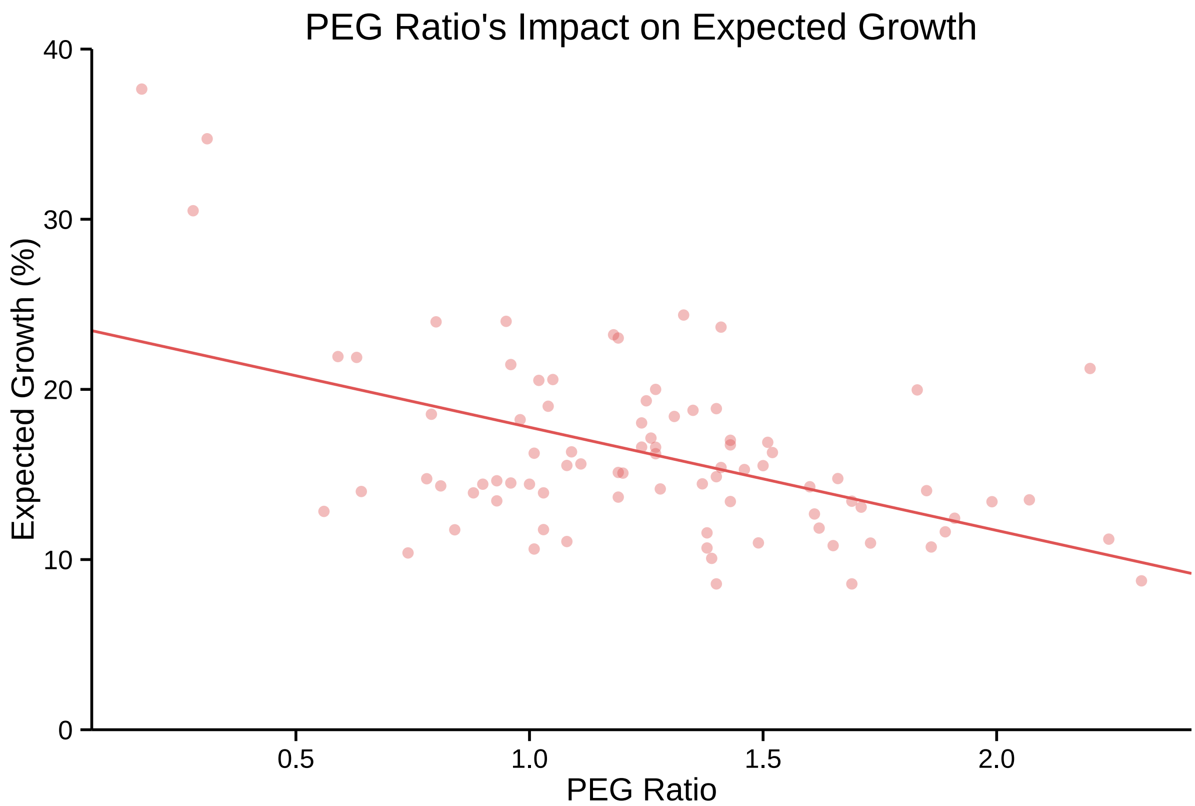
We found that when we account for the outliers, there's a slight positive multivariate correlation between expected growth, beta, and current PE. Continuing with the trend, we further investigated another related variables, the PEG Ratio. Our next regression checks for a linear relationship between PEG and growth, meanwhile retaining PE and unlevered beta in the regression to eliminate possible effects of confounding factors. We also included the regression of expected growth to PEG when discounting the effects of other variables. We additionally removed the outliers as we saw fit.

```
# Graph the scatterplot of PEG ratio's impact on expected growth and the
# regression line of expected growth to PEG ratio and the regression line
# of the multivariate regression of expected growth in the next 5
# years to average unlevered beta, the current PE ratio, and the PEG Ratio.
reg3 <- lm(Expected.Growth.Next.5.Years ~ Average.Unlevered.Beta +
               Current.PE + PEG.Ratio, data = clean_data)
reg4 <- lm(Expected.Growth.Next.5.Years ~ PEG.Ratio, data = clean_data)
ggplot(clean_data, aes(x = PEG.Ratio, y = Expected.Growth.Next.5.Years)) +
    geom_point(color = rgb(223, 84, 84, 100, maxColorValue = 255)) +
    xlab("PEG Ratio") + ylab("Expected Growth (%)") +
    geom_abline(aes(slope = unname(coef(reg3)["PEG.Ratio"]),
                    intercept = unname(coef(reg3)["(Intercept)"])),
                color = rgb(84, 233, 233, maxColorValue = 255)) +
    geom_abline(aes(slope = unname(coef(reg4)["PEG.Ratio"]),
                    intercept = unname(coef(reg4)["(Intercept)"])),
                color = rgb(223, 84, 84, maxColorValue = 255)) +
    ggtitle("PEG Ratio's Impact on Expected Growth") +
    coord_cartesian(ylim=c(0,40)) + theme_classic()
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



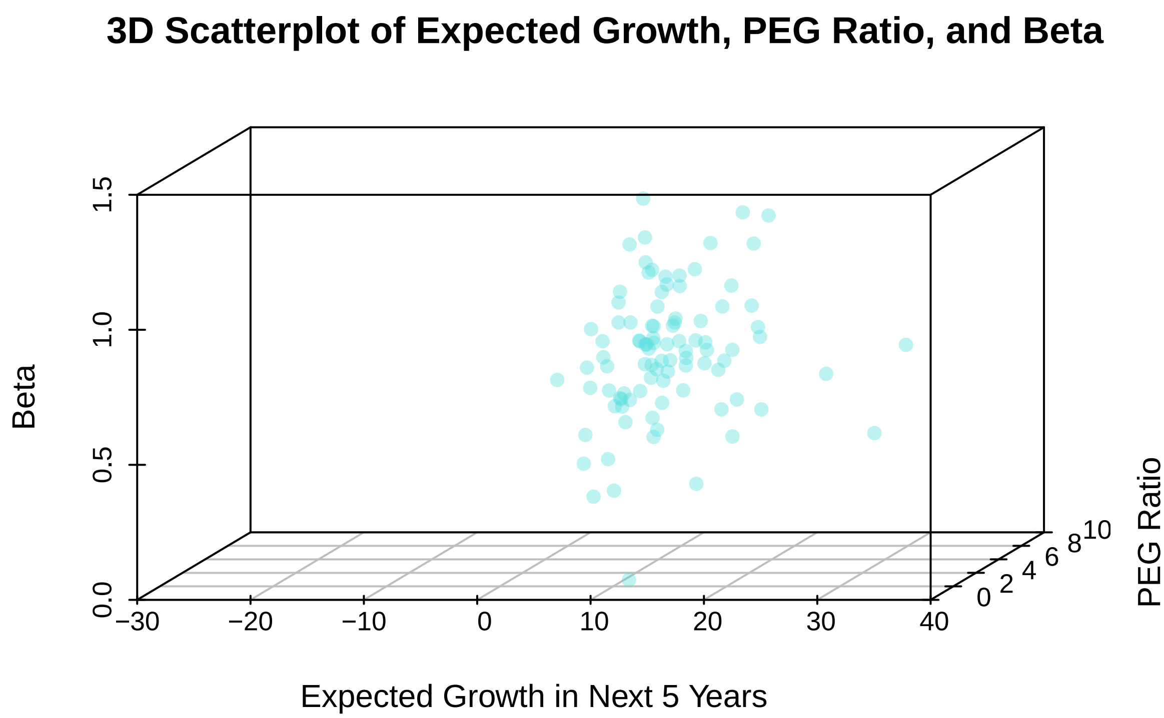

```
# Graph the scatterplot and regression of PEG ratio and expected growth
# accounting for outliers
peg_no_outliers <- clean_data$PEG.Ratio[indices_no_outliers]
dat9 <- data.frame(growth = growth_no_outliers, peg = peg_no_outliers)
dat9 <- subset(dat9, growth > 6.32 & peg < 2.5 & !is.na(peg))
reg4_no_outliers <- lm(growth ~ peg, data = dat9)
ggplot(dat9, aes(x = peg, y = growth)) + xlab("PEG Ratio") +
    ylab("Expected Growth (%)") + coord_cartesian(ylim=c(0,40)) +
    geom_point(color = rgb(223, 84, 84, 100, maxColorValue = 255)) +
    geom_abline(aes(slope = unname(coef(reg4_no_outliers)["peg"]),
                    intercept = unname(coef(reg4_no_outliers)["(Intercept)"])),
                color = rgb(223, 84, 84, maxColorValue = 255)) +
    ggtitle("PEG Ratio's Impact on Expected Growth") + theme_classic()
```



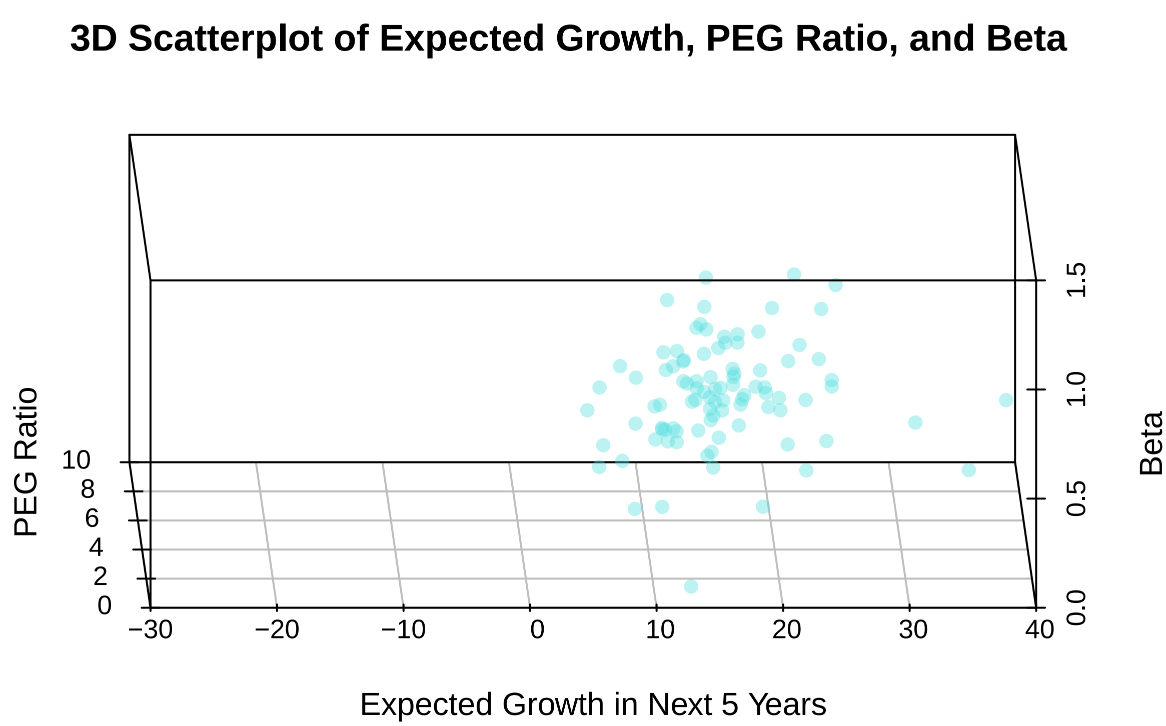
We find a statistical significance to the above coefficients, but the multivariate regression does not make a dramatic impact on best-fit line when compared to the normal regression. In the graph, the green line represents the multivariate regression while the red line represents the simple regression. Both regressions returned a negative slope for PEG and growth, meaning the industries had a relatively stable PE throughout.

We then visually examined a three-dimensional scatterplot of expected growth, PEG ratio, and average unlevered beta from multiple angles. This allows us to easily visualize the correlations between all of the variables.

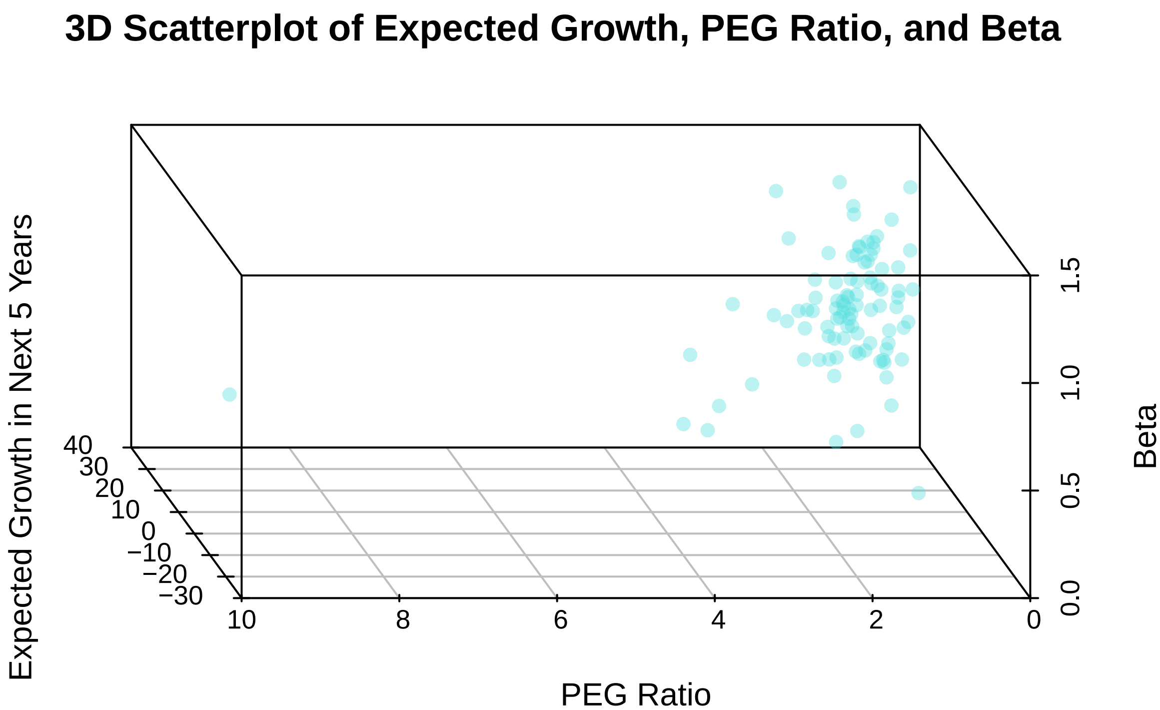
```
# Make 3D scatterplot of Expected Growth in the Next 5 Years, PEG Ratio,
# and Average Unlevered Beta, then examine the graph from multiple angles
scatterplot3d(x = clean_data$Expected.Growth.Next.5.Years,
              y = clean_data$PEG.Ratio,
              z = clean_data$Average.Unlevered.Beta,
              angle = 30, scale.y = .3, pch = 16,
              xlab = "Expected Growth in Next 5 Years",
              ylab = "PEG Ratio", zlab = "Beta",
              color = rgb(83, 223, 223, 100, maxColorValue = 255),
              main = "3D Scatterplot of Expected Growth, PEG Ratio, and Beta")
```




```r
scatterplot3d(x = clean_data$Expected.Growth.Next.5.Years,
              y = clean_data$PEG.Ratio,
              z = clean_data$Average.Unlevered.Beta,
              angle = 100, scale.y = .3, pch = 16,
              xlab = "Expected Growth in Next 5 Years",
              ylab = "PEG Ratio", zlab = "Beta",
              color = rgb(83, 223, 223, 100, maxColorValue = 255),
              main = "3D Scatterplot of Expected Growth, PEG Ratio, and Beta")
```



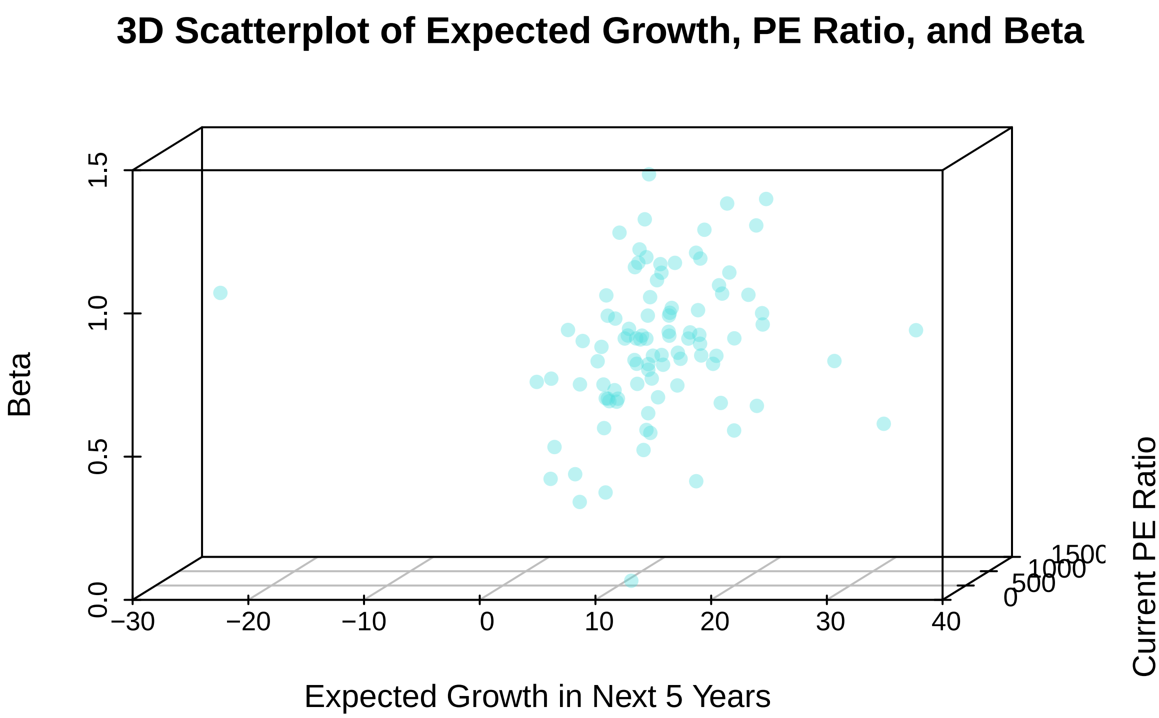
```r
scatterplot3d(x = clean_data$Expected.Growth.Next.5.Years,
              y = clean_data$PEG.Ratio,
              z = clean_data$Average.Unlevered.Beta,
              angle = 300, scale.y = .3, pch = 16,
              xlab = "Expected Growth in Next 5 Years",
              ylab = "PEG Ratio", zlab = "Beta",
              color = rgb(83, 223, 223, 100, maxColorValue = 255),
              main = "3D Scatterplot of Expected Growth, PEG Ratio, and Beta")
```



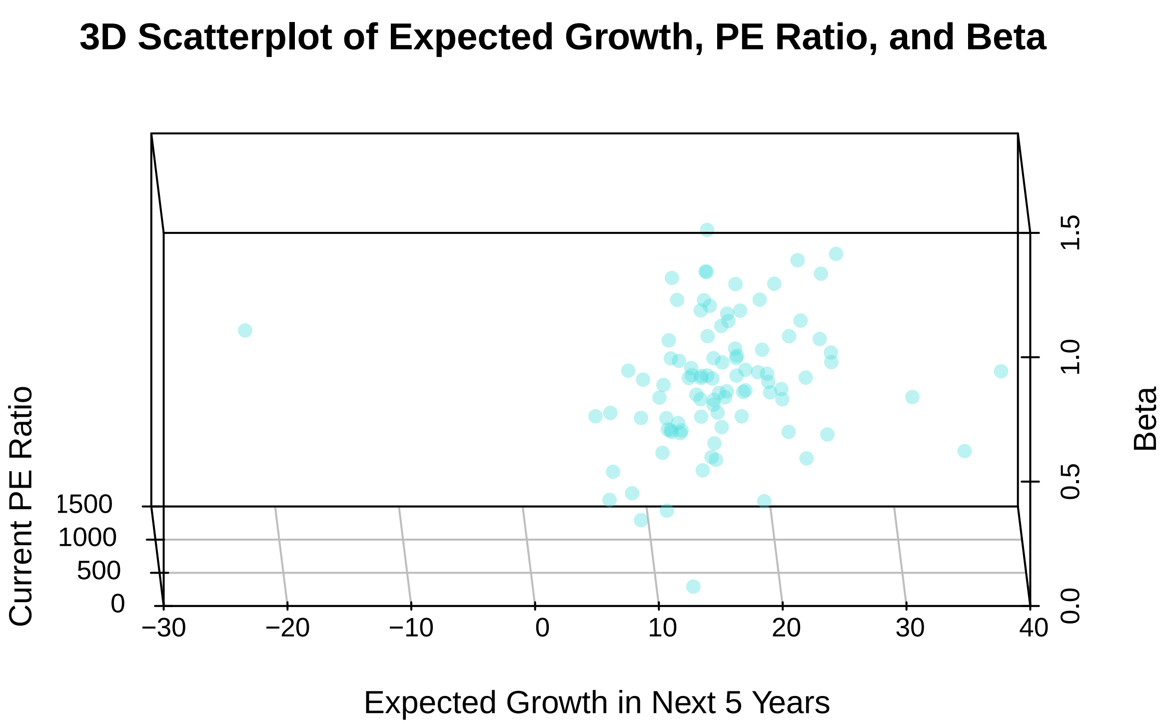
From our scatterplot, we see that low PEG ratios were associated with high levels of beta and expected growth. That is, expected growth and beta both have a negative impact on PEG ratios, since PEG ratio decreases as beta and expected growth increases. Conversely, higher PEG ratios are associated with lower levels of expected growth and beta and increases as expected growth and beta decreases.

We then examined the scatterplot of expected growth, current PE ratio, and average unlevered beta from multiple angles.

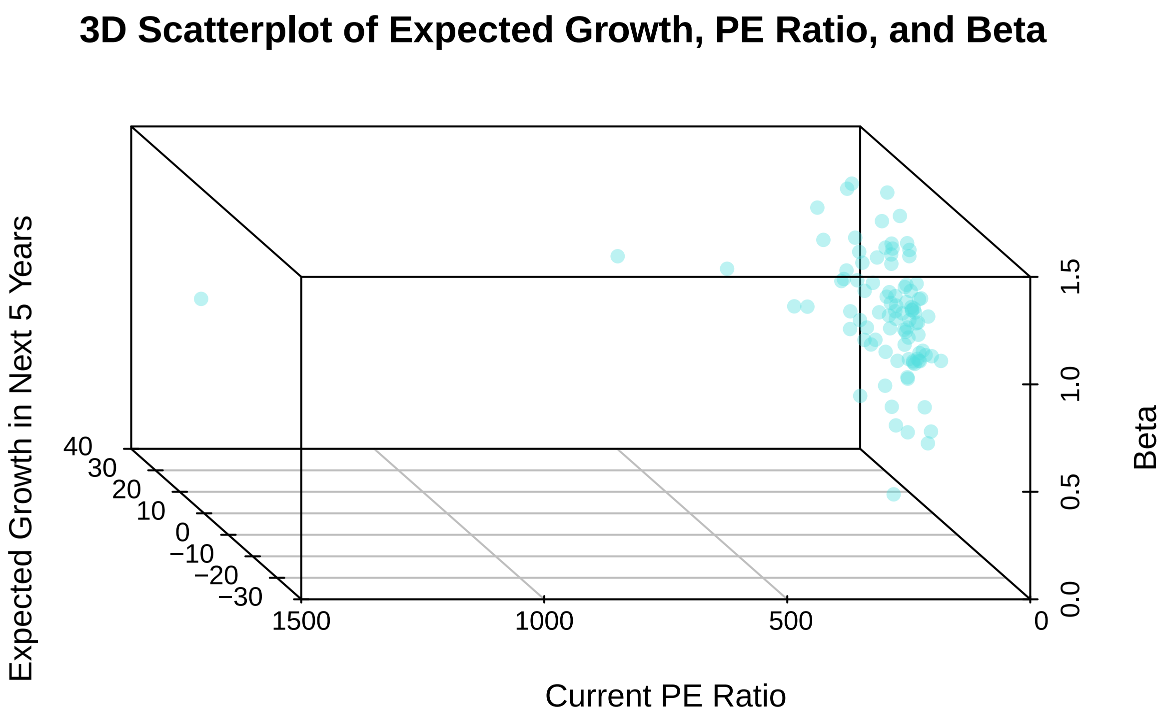
```r
# Make 3D scatterplot of Expected Growth in the Next 5 Years, Current PE,
# and Average Unlevered Beta, then examine the graph from multiple angles
scatterplot3d(x = clean_data$Expected.Growth.Next.5.Years,
              y = clean_data$Current.PE, z = clean_data$Average.Unlevered.Beta,
              angle = 30, scale.y = .3, pch = 16,
              xlab = "Expected Growth in Next 5 Years",
              ylab = "Current PE Ratio", zlab = "Beta",
              color = rgb(83, 223, 223, 100, maxColorValue = 255),
              main = "3D Scatterplot of Expected Growth, PE Ratio, and Beta")
```




```
scatterplot3d(x = clean_data$Expected.Growth.Next.5.Years,
              y = clean_data$Current.PE, z = clean_data$Average.Unlevered.Beta,
              angle = 100, scale.y = .3, pch = 16,
              xlab = "Expected Growth in Next 5 Years",
              ylab = "Current PE Ratio", zlab = "Beta",
              color = rgb(83, 223, 223, 100, maxColorValue = 255),
              main = "3D Scatterplot of Expected Growth, PE Ratio, and Beta")
```



```r
scatterplot3d(x = clean_data$Expected.Growth.Next.5.Years,
              y = clean_data$Current.PE, z = clean_data$Average.Unlevered.Beta,
              angle = 300, scale.y = .3, pch = 16,
              xlab = "Expected Growth in Next 5 Years",
              ylab = "Current PE Ratio", zlab = "Beta",
              color = rgb(83, 223, 223, 100, maxColorValue = 255),
              main = "3D Scatterplot of Expected Growth, PE Ratio, and Beta")
```



We found that low current PE ratios were associated with high levels of beta and expected growth. Expected growth and beta both have a negative impact on current PE, while current PE ratio increases as expected growth and beta decreases.

Findings & Conclusion

Through our analysis, we tried to answer our hypothesis through our guiding questions, which we answer below.

- We found from our analysis that PE is negatively correlated with beta. That is, as beta increases, PE decreases. Thus, there is a negative correlation between beta and the payout ratio.
- Industries with higher beta also have higher growth rates, as we saw in our findings. Thus, industries such as technology – software and online retail – have higher growth but are less stable than industries such as finance, which also have a lower beta.

Our findings state that industries with higher beta also has higher growth. However, because higher beta is negatively correlated with payout ratio, then industries with higher beta also has a low PE. Thus, for the technology industry with a higher beta, it also has high growth, which also leads the industry to have a low PE value.